Inflammatory biomarker genomics

Prins, Bram Peter

Publication date:
2016

Inflammatory biomarker genomics

Bram Peter Prins

Bram Peter Prins
Inflammatory biomarker genomics.

The research in this thesis was supported by a Netherlands Organization for Scientific Research (NWO) VENI grant (ZonMW grant 016.096.121). Printing of this thesis was financially supported by Rijksuniversiteit Groningen, University Medical Center Groningen.

ISBN: 978-90-367-9033-8 (e-book)
ISBN: 978-90-367-9037-6 (print)

Cover and layout design by Claudia González Arévalo
www.claudiagonzaleza.com

Printed by NetzoDruk Groningen B.V.

rijksuniversiteit
groningen

Inflammatory biomarker genomics

From discovery to causality

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

woensdag 7 september 2016 om 12.45 uur

door

Bram Peter Prins
geboren op 29 april 1979
te Dirksland

Promotor
Prof. dr. H. Snieder

Copromotor
Dr. B.Z. Alizadeh

Beoordelingscommissie
Prof. dr. H.M. Boezen
Prof. dr. M.A. Swertz
Prof. dr. A.P. Morris

Paranimfs:
Yvonne Chen
Miriam Prins

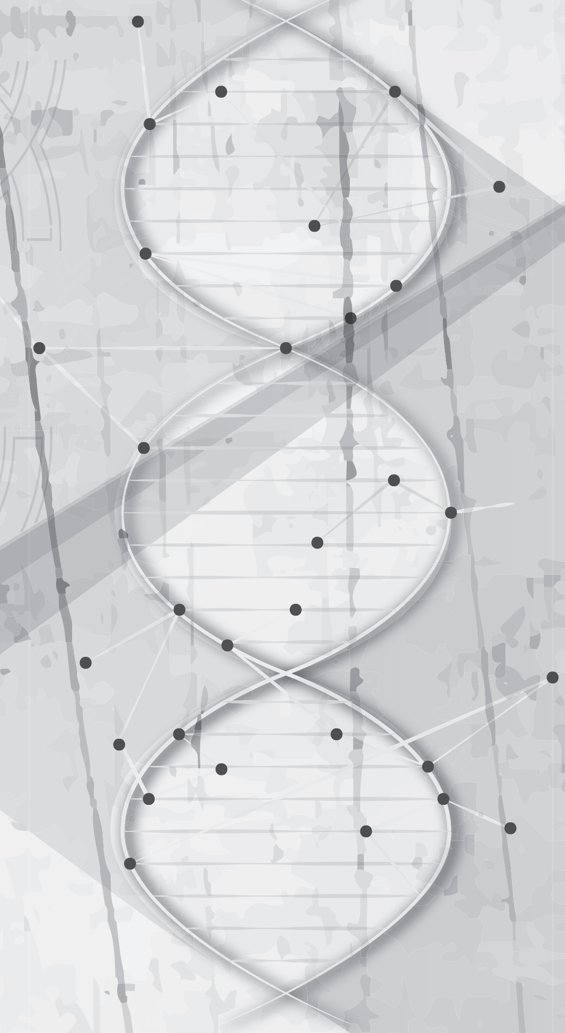
To my family

致我的家人

SECTION I :
GWAS AND INFLAMMATORY MARKER GENETICS

SECTION II:
INTEGRATIVE POST-GWAS ANALYSES AND SYSTEMS GENETICS

SECTION III :
GENERAL SYNTHESIS, BROADER AND FUTURE PERSPECTIVES



1

Preface and structure of the thesis

PREFACE

Inflammation is the body's normal response against tissue injury or infection, instigated by the immune system and characterised by vascular permeability, increased blood flow along with the build up of fluid and an increase of white blood cells. It is a complex but harmonic process that involves various types of immune cells, each with specialised functions, and orchestrated through molecular mediators known as inflammatory biomarkers.[1] One important class of inflammatory biomarkers are the cytokines, which are molecular messengers that, in concert with specific cytokine inhibitors and soluble cytokine receptors, facilitate signalling between immune cells to coordinate the immune response[2]. Classically, two broad classes of cytokines can be distinguished[3]. One class are the so-called pro-inflammatory cytokines, which promote and propagate inflammation, whereas the other group, anti-inflammatory cytokines, inhibit inflammation. Under normal circumstances, cytokines control the immune response so as to repair and restore normal tissue function, whereby a careful balance between pro- and anti-inflammatory cytokines is maintained. In certain circumstances, an imbalance of these two classes of cytokines occurs, resulting in an exaggerated immune response, coincided by excessive levels of inflammatory markers, which may have various negative consequences, the most important being tissue damage[4]. Elevated levels of certain cytokines are hallmarks for various types of diseases[5], and sometimes have prognostic value[6]. In certain cases they are even directly implicated in pathogenesis[7], and therefore considered as therapeutic targets[5].

Besides being influenced by environmental factors, serum levels of inflammatory markers are partially genetically determined, that is, they are heritable[8]. Linking to the above, this also implies that those diseases associated with elevated inflammatory markers have a partly genetically regulated inflammatory component, which may contribute to their genetic susceptibility.

Prior to the advent of genome-wide association studies (GWAS), identification of genetic polymorphisms involved in complex polygenic traits, such as inflammatory marker levels, depended on rather circumstantial evidence for certain genomic regions investigated in candidate gene studies with often non-replicable results[9]. Linkage studies provided a more systematic way to query the genome, but were hampered by a rather low granularity;

results would not necessarily apply to the general population as linkage studies depend on familial relatedness, and proved to be more suitable for simple Mendelian models[10,11]. Two main developments boosted the identification of genetic variation underlying complex polygenic traits. Firstly, the Human Genome Project[12,13], launched in 1990 and completed in 2003, which established a human DNA sequence reference, and for the first time systematically catalogued genes in a physical and functional manner.

The second important development was the Hapmap Project, it's first release being in 2005[14], which established a catalogue of single nucleotide polymorphisms (SNPs), which are the most abundant human genetic variants that vary between individuals on a single DNA letter. A key finding from this effort was the identification of haplotype blocks: blocks of DNA along a chromosome that have low recombination rates, characterized by relatively few haplotypes. Haplotype blocks are groups of SNPs in high linkage disequilibrium (LD), which is formally defined as the non-random association of alleles at different loci[15]. In other words, it is the extent to which an allele from one SNP is observed together with that of another as a combination. The more often this particular combination is seen, the higher the degree of linkage disequilibrium, that is, the occurrence of these alleles is correlated.

That makes it possible, when SNPs are highly correlated (exhibiting a high linkage disequilibrium), to predict for the allele for one SNP, when knowing the particular allele for another. In other words, one SNP 'tags' the other. Therefore we just need to analyse a subset of tagging SNPs in order to be able to conclude something about the majority of tagged SNPs. This makes genotyping the large numbers of individuals required for genetic analyses financially feasible. Moreover, by making use of the same tag SNP phenomenon, unobserved variants can be inferred from known variants in the same haplotype block, meaning these variants can be imputed once haplotypes are known[16], for example, from reference sequence population panels such as those from Hapmap[14] or 1000 Genomes[17].

The GWAS approach, whereby hundreds of thousands of markers across the genome could be analysed systematically became a standard approach to unravel the genetic architecture of a multitude of traits and diseases. It should be noted here, that many sources state the first GWAS experiment was performed by Haines et.al. in 2005 on age-related macular degeneration[18], which is incorrect. In fact, it was Ozaki and colleagues that pioneered what now is referred to as a GWAS on myocardial infarction as early as in 2002[19], the same year the Hapmap project started. It became clear early on however, that having found that the genetic variation underpinning many complex polygenic traits only explained a small part of the total genetic variance (i.e., heritability), for various reasons[20–22]. One obvious reason is that there is additional genetic variation that was as of yet undiscovered, because of insufficient statistical power to detect variants with very small effects[23]. Statistical power can be improved by increasing sample sizes amongst other things[24]. This spurred the formation of large GWAS collaborations in the form of consortia, where GWAS results from individual efforts were combined in GWAS meta-analyses[25–27]. This approach quickly bore its fruits, demonstrated by the thousands of replicable genetic loci that have been identified for hundreds of traits and diseases[28]. This holds true for investigations involving genetic determinants of inflammatory markers levels as well[29].

Even so, having an ever-expanding atlas of genetic loci underpinning complex diseases and traits[30], does not imply that we understand how these loci and the variants that they harbour contribute to the phenotype of interest. To this end, an entire new branch of bioinformatics has naturally arisen, with the specific aim to design algorithms and analysis strategies to understand the mechanisms through which genetic variants exert their effects on a phenotype, in this context collectively known as post-GWAS analyses[31]. As of now, performing follow-up analyses are a must when performing a GWAS meta-analysis, with the aim to understand the phenotype of interest from the variant level, intermediate levels such as mRNA expression, epigenetic effects and protein levels, and finally to understand the collective effects in molecular pathways and interaction networks by integrating these in so-called integrative multi-omics analyses[32]. One step further is to understand what are the main tissues in which these variants have their main physiological consequences, and finally through wet-lab experiments, be it in cellular or animals models, confirm modes of action and mechanisms[33].

Coming back to inflammatory markers, various loci have been established through GWAS for a number of major players, such as C-reactive protein (CRP)[29], Interleukin 1 receptor antagonist (IL1-RA) and Interleukin-18 (IL-18)[34], soluble Interleukin-6 Receptor (sIL-6R)[35] just to mention a few. Nevertheless, for most of these cytokines, with the exception of CRP, very few loci have been established, explaining only a small fraction of the estimated heritabilities.

This thesis has three major aims: (i) to identify novel genetic loci for major inflammatory markers, and; (ii) to understand functional mechanisms underlying these genetic loci, and; (iii) to understand the causal impact of inflammatory markers on disease. The thesis is divided in three sections in which I consecutively identify novel loci for inflammatory markers, followed by fine-mapping and assessment of functional mechanisms and causal involvement in disease.

Finally I will end with an in-depth discussion, integrating and placing major findings in a broader context and presenting future perspectives and recommendations for further research.

THESIS OUTLINE

Section I : GWAS and inflammatory marker genetics

In Chapter 2, we present a software suite that we developed, QCGWAS, that automates the quality control of genome-wide association result files, thereby facilitating the rapid generation of high-quality input files for meta-analysis of genome-wide association studies.

In Chapter 3, I describe the first ever GWAS meta-analysis for circulating levels of Tumor Necrosis Factor (TNF), encompassing more than 30,000 individuals of European descent. This effort led to the establishment of 3 novel genetic loci associated with circulating levels of TNF, linking these to previously unsuspected biological mechanisms involved.

Chapter 4 entails a similar exercise, whereby through the establishment of large meta-GWAS consortium for Interleukin 6 (IL6) encompassing over 60,000 individuals from European descent, we identify 3 novel loci associated with this inflammatory marker.

In Chapter 5 we describe the application of a novel GWAS meta-analysis method, MANTRA, to identify loci associated with serum levels of albumin and total protein in European and Asian populations. In total, we identify 9 novel loci associated with these traits. By capitalising on the different LD structures between these populations, we were able to define narrow regions of association that with 99.9% probability contain variants causal to changes in circulating levels for these biomarkers. Extensive follow-up bioinformatics analyses highlight a role for immune-response signalling, ribosomal functioning, protein translation and proteasomal protein degradation underlying these phenotypes.

Section II : Integrative post-GWAS analyses and systems genetics

In Chapter 6 we set out to disentangle the molecular mechanisms through which genetic variants associated with serum CRP levels exert their effects. Following and extending the same analytical post-GWAS approach as in Chapter 6, we firstly identified moderate to high LD variants in 1000 Genomes sequence data, as well as in expression quantitative loci (eQTLs), followed by mapping the entire collection of involved genetic variants to their nearest genes. Next, we used the collection of identified genes as input in a network analyses and subsequently conducted a functional enrichment analysis, that firstly confirmed an overlap between CRP and lipid biology, and secondly identified a previously unknown major role for interferons in the metabolism of CRP.

Chapter 7 entails the determination of causal involvement of CRP in a panel of 32 different traits and diseases in 5 broad outcome classes, being auto-immune, cardiovascular, metabolic, neuro-degenerative and psychiatric diseases. The investigated outcomes were selected for being accompanied by or associated with elevated levels of circulating CRP. By using genetic variants as so-called instrumental variables, modelling serum CRP levels in a Mendelian Randomisation analysis framework, using GWAS summary statistics only, we found that CRP is not causally contributing to disease risk in the majority of the outcomes that were considered.

One surprising exception was for schizophrenia, for which we found a protective, potentially causal, effect of CRP, which we confirmed using individual-level data.

In Chapter 8 we review and evaluate the success of GWAS meta-analyses and their ability to contribute towards the unexplained heritability for coronary artery disease (CAD), which is a complex disease in which inflammation plays a leading role through atherosclerosis.

We finally argue that simply solving the missing heritability problem is merely just one step towards being able to understand the genetic architecture of a complex polygenic disease, and propose a systems genetics approach as the way forward to understand the etiology of CAD.

Section III : General synthesis, broader and future perspectives

In Chapter 9 I synthesize the findings in this thesis and discuss these from multiple perspectives; from clinical interpretation and impact to statistical genetics. I conclude with some future perspectives and give potentially useful recommendations for future research.

REFERENCES

1. Medzhitov, R. Origin and physiological roles of inflammation. Nature 454, 428–435 (2008).
2. Kelso, A. Cytokines: Principles and prospects. Immunol Cell Biol 76, 300–317 (1998).
3. Cavaillon, J. M. Pro- versus anti-inflammatory cytokines: myth or reality. Cell. Mol. Biol. (Noisy-le-grand) 47, 695–702 (2001).
4. Serhan, C. N., Ward, P. A. & Gilroy, D. W. Fundamentals of Inflammation. (Cambridge University Press, 2010).
5. Kopf, M., Bachmann, M. F. & Marsland, B. J. Averting inflammation by targeting the cytokine environment. Nat Rev Drug Discov 9, 703–718 (2010).
6. Pearson, T. A. et al. Markers of Inflammation and Cardiovascular Disease Application to Clinical and Public Health Practice: A Statement for Healthcare Professionals From the Centers for Disease Control and Prevention and the American Heart Association. Circulation 107, 499–511 (2003).
7. McInnes, I. B. & Schett, G. Cytokines in the pathogenesis of rheumatoid arthritis. Nat Rev Immunol 7, 429–442 (2007).
8. Sas, A. A. et al. The age-dependency of genetic and environmental influences on serum cytokine levels: a twin study. Cytokine 60, 108–113 (2012).
9. Chanock, S. J. et al. Replicating genotype–phenotype associations. Nature 447, 655–660 (2007).
10. Stranger, B. E., Stahl, E. A. & Raj, T. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. Genetics 187, 367–383 (2011).
11. Londin, E., Yadav, P., Surrey, S., Kricka, L. J. & Fortina, P. Use of linkage analysis, genome-wide association studies, and next-generation sequencing in the identification of disease-causing mutations. Methods Mol. Biol. 1015, 127–146 (2013).
12. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
13. A list of authors and their affiliations appears in the Supplementary Information et al. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945 (2004).
14. The International HapMap Consortium. A haplotype map of the human genome. Nature 437, 1299–1320 (2005).
15. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9, 477–485 (2008).
16. Scheet, P. & Stephens, M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am J Hum Genet 78, 629–644 (2006).
17. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).
18. Haines, J. L. et al. Complement factor H variant increases the risk of age-related macular degeneration. Science 308, 419–421 (2005).
19. Ozaki, K. et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat. Genet. 32, 650–654 (2002).
20. Maher, B. Personal genomes: The case of the missing heritability. Nature 456, 18–21 (2008).

21. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).

22. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11, 446–450 (2010).

23. McCarthy, M. I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9, 356–369 (2008).

24. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 15, 335–346 (2014).

25. GAIN Collaborative Research Group et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat. Genet. 39, 1045–1051 (2007).

26. Psaty, B. M. et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2, 73–80 (2009).

27. Bennett, S. N. et al. Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. Genet. Epidemiol. 35, 159–173 (2011).

28. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. Am J Hum Genet 90, 7–24 (2012).

29. Dehghan, A. et al. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123, 731–738 (2011).

30. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42, D1001–1006 (2014).

31. Freedman, M. L. et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 43, 513–518 (2011).

32. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. 16, 85–97 (2015).

33. Lehner, B. Genotype to phenotype: lessons from model organisms for human genetics. Nat Rev Genet 14, 168–178 (2013).

34. Matteini, A. M. et al. Novel gene variants predict serum levels of the cytokines IL-18 and IL-1ra in older adults. Cytokine 65, 10–16 (2014).

35. Melzer, D. et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 4, e1000072 (2008).

SECTION I:
GWAS AND INFLAMMATORY MARKER GENETICS

2

QCGWAS: A flexible R package for automated quality control of genome-wide association results

Bioinformatics. 2014 Apr 15;30(8):1185-1186

Peter J. van der Most, Ahmad Vaez*, Bram P. Prins*, M. Loretto Munoz, Harold Snieder, Behrooz Z. Alizadeh and Ilja M. Nolte.

*Equal contribution

ABSTRACT

Summary: QCGWAS is an R package that automates the quality control of genome-wide association result files. Its main purpose is to facilitate the quality control of a large number of such files before meta-analysis. Alternatively, it can be used by individual cohorts to check their own result files. QCGWAS is flexible and has a wide range of options, allowing rapid generation of high-quality input files for meta-analysis of genome-wide association studies.

Availability: http://cran.r-project.org/web/packages/QCGWAS

INTRODUCTION

The number of consortia aiming to identify genes for complex traits through meta-analysis of genome-wide association studies (GWAS) has mushroomed in the past 6 years. The advantage of this strategy is that large sample sizes can be reached, allowing detection of genetic variants with small effects. A downside is the lack of unified quality control (QC) on the GWAS analyses of the individual cohorts, as each cohort will typically perform their own analysis according to a standard analysis plan and share only summary statistics. GWAS result files are prone to errors due to the vast amount of data they contain and the different manner in which these data are generated by individual cohorts. Before combining data from individual studies in a meta-analysis, it is important to ensure that all data included are valid, of high quality and compatible between cohorts to reduce both the false-positive and the false-negative findings[1]. Because GWAS result files usually contain a standard set of variables, it is feasible to automate the QC of these files, thereby gaining speed, reliability, flexibility and the possibility to perform more elaborate checks.

To our knowledge, the only other software package currently available for QC of GWAS result files is GWAtoolbox[2]. However, GWAtoolbox does not produce cleaned results files, is less flexible regarding file format and uses a restrictive format for the QC log. This makes it less suited for processing (and comparing) large numbers of files in preparation of a meta-analysis. It also does not check allele information or allow for the retesting of individual QC steps. To address these shortcomings, we developed QCGWAS with the aim to automate QC and allow rapid generation of high-quality input files for GWAS meta-analyses.

APPROACH

Implementation

QCGWAS is built as a package for R[3]. The R platform was chosen because it is operating system-independent, commonly used, open source, can handle large datasets and is flexible regarding input file format. QCGWAS requires R version 3.0.1 or later (64-bit recommended) and can be downloaded from the Comprehensive R Archive Network Web site (http://cran.r-project.org).

Usage

The main QC by QCGWAS is executed by the QC_series(. . .) command. This function requires a minimum of two parameters: a list of filenames of GWAS result files and a translation table for the file headers. All other parameters are optional, allowing for a flexible and user-customized QC.

Approach

A standard QC consists of six steps (Fig. 1):

STAGE 1: a GWAS result file is inspected for missing and invalid data. Duplicated single nucleotide polymorphisms (SNPs) and SNPs lacking crucial variables are removed.

STAGE 2: alleles and strand information are checked and fixed by matching it to a given reference (e.g. HapMap). The SNPs can be removed when their alleles or allele frequencies do not match the reference. This harmonizes the alleles across result files. Next, it correlates the reported allele frequencies for all SNPs to those from the reference set and generates scatter plots to show deviations (Supplementary Fig. S1).

STAGE 3: QC plots are generated (see Supplementary Fig. S2–S4). These include histograms of the distribution of SNP quality parameters (allele frequencies, Hardy-Weinberg equilibrium P-values, call rates and imputation quality), a Manhattan plot and a series of Quantile-Quantile (QQ) plots filtered for SNP quality.

STAGE 4: various QC statistics are calculated, of which the most important are the genomic-control lambda[4], Visscher's statistic[5] to determine whether the standard errors are in line with the sample size reported, the skewness and kurtosis of the effect-size distribution and the correlation between the reported P-values and those calculated from the effect size and standard error.

STAGE 5: the cleaned GWAS result file is saved and extensive QC information is written to a log file. The cleaned file can be saved in different formats, ensuring compatibility for immediate meta-analysis by GWAMA[6], META[7], MetABEL[8], METAL[9] or PLINK[10].

STAGE 6: several between-study checks are performed, including a comparison of

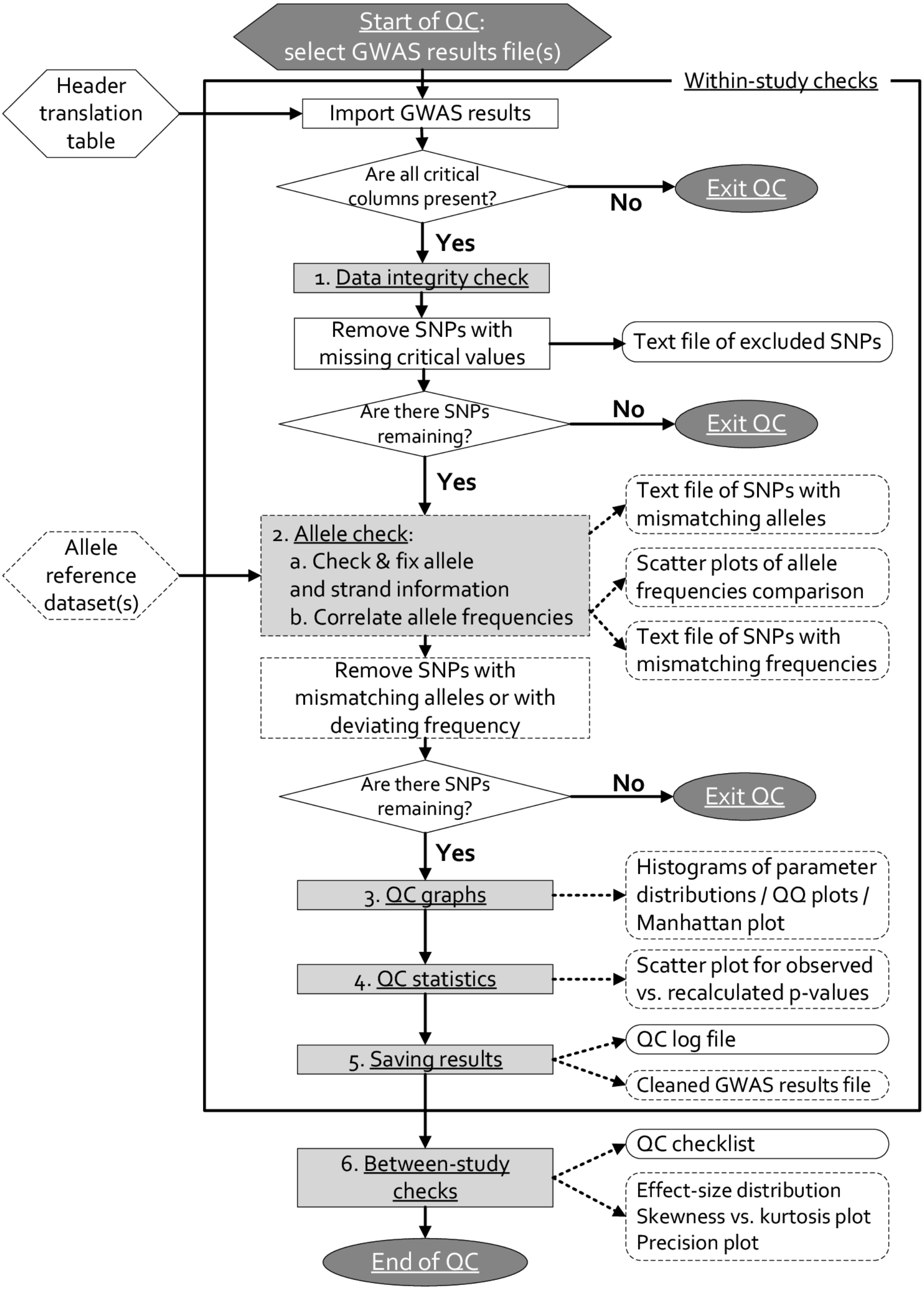

Start of QC:
select GWAS results file(s)
Within-study checks
Header translation table
Import GWAS results
Are all critical columns present?
No
Exit QC
Yes
1. Data integrity check
Remove SNPs with missing critical values
Text file of excluded SNPs
Are there SNPs remaining?
No
Exit QC
Yes
Text file of SNPs with mismatching alleles
Allele reference dataset(s)
2. Allele check:
a. Check & fix allele and strand information
b. Correlate allele frequencies
Scatter plots of allele frequencies comparison
Text file of SNPs with mismatching frequencies
Remove SNPs with mismatching alleles or with deviating frequency
Are there SNPs remaining?
No
Exit QC
Yes
3. QC graphs
Histograms of parameter distributions / QQ plots / Manhattan plot
4. QC statistics
Scatter plot for observed vs. recalculated p-values
5. Saving results
QC log file
Cleaned GWAS results file
6. Between-study checks
QC checklist
Effect-size distribution
Skewness vs. kurtosis plot
Precision plot
End of QC

Figure 1. Flow diagram of the six steps (marked by light grey-shaded rectangles) comprising the default QC performed by QCGWAS. Input files are indicated by hexagons and the created output files by rounded rectangles. Dashed lines indicate that the check is optional.

skewness and kurtosis, of sample sizes and standard errors and of effect-size range to identify incorrect units and/or trait transformations (Supplementary Fig. S5). A checklist of QC statistics is also created.

Each of the steps of the QC can be enabled or disabled by the user, allowing for a flexible QC pipeline, and quick retests of particular steps. Finally, independent functions are provided for the creation of histograms or QQ plots using combinations of filter parameters and regional association plots.

Performance

On a Windows 7 computer with 2.4 GHz and 48GB RAM, a QC of a HapMap-imputed GWAS result file (2.5 million SNPs) takes between 5 and 15 min/file. Memory usage is between 2 and 3GB, depending on the number of graphs to be created. Sequence-imputed results files, such as 1000 Genomes-based data[11] take ~40 min and 20GB of RAM.

CONCLUSION

QCGWAS is a flexible and comprehensive package for automated QC of GWAS result files. It can handle a large number of files within reasonable time and is therefore particularly useful for a centralized QC preceding a GWAS meta-analysis. It can also be used by individual cohorts to inspect the quality of their results. Currently it is geared toward quantitative traits, but case-control results can also be used with proper transformations. Future versions of the package are under development to accommodate non- SNP variants, such as used in sequence-based GWAS data.

Acknowledgements

The authors would like to acknowledge Josée Dupuis for constructive discussions and for sharing her QC procedure and scripts at an early stage. They are also grateful to Nicola Barban and Jornt Mandemakers for their useful feedback on the use of QCGWAS. Conflict of Interest: none declared.

Supplementary material

Supplementary Material is available at Bioinformatics online.

REFERENCES

1. de Bakker, P. I. W. et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. 17, R122–128 (2008).
2. Fuchsberger, C., Taliun, D., Pramstaller, P. P., Pattaro, C. & CKDGen consortium. GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. Bioinformatics 28, 444–445 (2012).
3. Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. (2005). at <http://www.R-project.org>
4. Devlin, B. & Roeder, K. Genomic control for association studies. Biometrics 55, 997–1004 (1999).
5. Yang, J. et al. FTO genotype is associated with phenotypic variability of body mass index. Nature 490, 267–272 (2012).
6. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11, 288 (2010).
7. Liu, J. Z. et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat. Genet. 42, 436–440 (2010).
8. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294–1296 (2007).
9. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).
10. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
11. 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).

SECTION I :
GWAS AND INFLAMMATORY MARKER GENETICS

3

Meta-GWAS analysis of TNF-α levels in over 30,000 individuals reveals novel genetic loci

On behalf of the TNF-α Meta-GWAS Consortium:

Bram P. Prins, co-authors of the TNF-α Meta-GWAS Consortium, co-authors of the CHARGE Inflammation Working Group, Harold Snieder and Behrooz Z. Alizadeh

ABSTRACT

Tumor Necrosis Factor alpha (TNF-α) is a pro-inflammatory cytokine that is involved in a wide range of biological processes, and can be synthesized by multiple cell-types. Its involvement and elevated serum levels in a phletora of diseases ranging from rheumatoid arthritis to cardiovascular diseases, has ensured it has become a primary drug-target and warranted a continuous interest in the development of TNF-α inhibiting drugs. Even though previous studies investigating the genetic architecture of variation of TNF-α have been unsuccessful, up to 39% of the variation is estimated to be heritable, leaving much ground for uncovering genetic determinants of TNF-α levels. We have formed a consortium of 26 cohorts, enabling substantial gains in statistical power to identify genetic variants underpinning levels of TNF-α by combining genome-wide genetic association data encompassing up to 30,912 individuals in a meta-analysis. We identified 3 independent novel loci at chromosome 6p21 (rs2857602, $p=3.30x10^{-12}$), 12q24 (rs10744774, $p=6.94x10^{-12}$) and 15q21 (rs7182229, $p=1.07x10^{-9}$), harbouring genes with for cytokines and regulators of cytokines (6p21 / *LTA* and 12q24 / *SH2B3*) but also those involved in lipid metabolism (15q21 / *LIPC*) and tumor suppression (12q24 / *SH2B3*). Variants harboured within or nearby these genes affect a wide range of other traits and diseases, including common cardiovascular traits and diseases including red blood cell traits, coronary heart disease, diastolic and systolic blood pressure and confirm roles of TNF-α in auto-immune diseases such as rheumatoid arthritis, type 1 diabetes and celiac disease. Our work constitutes the first genome-wide association study discovering and replicating variants significantly associated with TNF-α, and offers novel biological insights in their involvement in a range of traits and diseases.

INTRODUCTION

Tumor Necrosis Factor alpha (TNF-α) is a pro-inflammatory cytokine that is involved in a wide range of biological processes, including the immune response, cell metabolism, differentiation, proliferation and apoptosis[1]. It is predominantly synthesized by macrophages, though other cell types are known to produce *TNF* as well. After being synthesized in the form of a 26-kDa transmembrane monomer (mTNF)[2] ,the TNF-α converting enzyme (TACE) subsequently cleaves the monomer to a 17-kDa soluble TNF-α molecule (sTNF)[3]. In soluble form, it exerts its biological activities through binding to TNF-α receptors 1 and 2 (TNF-R1 and -R2) on targeted cells. Given its primary role as a key regulator of the immune response and associated biological processes, significant changes in serum TNF-α levels may lead to an impaired immune response and exaggerated inflammation and subsequent cellular damage[4]. It is therefore not surprising that elevated serum levels of TNF-α are associated with a variety of pathological conditions and chronic diseases including autoimmune diseases such as rheumatoid arthritis[5] (RA) and inflammatory bowel disease (IBD)[6], infectious diseases of various origins (viral, bacterial or parasitic)[7], cardiovascular conditions such as heart failure and myocardial infarction[8,9], neurological diseases such as Alzheimer's disease[10] and psychiatric disorders such as depression[11]. Due to its functional properties TNF-α is a primary drug target of interest, and several biologic anti-TNF-α agents such as infliximab and etanercept are used in clinical treatments of, for example, RA and IBD, though with a range of side-effects[6], of which opportunistic infections are a well-known example[12,13].

Variation in baseline serum levels of TNF-α has a genetic component with heritability estimates of twin studies ranging from 0.17 to 0.39[14,15]. A number of studies have sought to identify genetic variants that determine serum levels TNF-α of which many were candidate gene studies, focussing on a handful of variants in the telomeric class III region of the HLA complex that harbours the *TNF* and *LTA* genes[16-19]. To date, genome-wide searches for variants regulating serum TNF-α levels in healthy individuals have not been successful. Two previous genome-wide association studies (GWAS), both investigating genetic associations with a large panel of biomarkers did initially identify variants associated with serum levels of TNF-α in the discovery phases, but these results could not be replicated[20,21], likely owing to relatively small samples sizes of discovery/ replication of 1,187/1,768 [20] or 1,167/708 subjects[21], respectively.

To identify and refine our understanding of the genetic factors that determine TNF-α serum levels, we performed the first large-scale meta-analysis of GWASs with a combined sample size of 30,912 subjects, consisting of a discovery of 23,141 individuals and subsequent replication of suggestively associated variants in an additional independent samples of 7,771 persons, offering a substantial increase in power compared to previous studies.

METHODS

Discovery Study

Study population

The discovery stage included individuals from 16 cohorts of European ancestry including Cohorte Lausannoise (CoLaus, 5,320), the Framingham Heart Study (FHS, N=2,315), FINRISK (N=213), Genetics of Lipid Lowering Drugs and Diet Network (GOLDN, N=816), Health 2000 (H2000, N=378), Health ABC (HABC, N=1,577), the Helsinki Birth Cohort Study (HBCS, N=1,688), the Lothian Birth Cohort of 1921 (LBC, N=169), the Ludwigshafen Risk and Cardiovascular Health study (LURIC, N=156), the Osteoporotic Fractures in Men study in the United States (MrOS US, N=756), the Netherlands Study of Depression and Anxiety (NESDA, N=1,895), the Netherlands Twin Registry (NTR, N=3,147), the Prospective Investigation of the Vasculature in Uppsala Seniors study (PIVUS, N=934), the Rotterdam Study (RS, N=833), TwinsUK (N=980) and the Young Finns Study (YFS, N=1,978). Only population based cohorts or healthy controls from case-control studies were included in the analyses.

TNF serum level measurements

Each study typically collected venous blood samples from their participants frozen as either serum or plasma and stored below -80°C until the time of measurement. Serum or plasma levels of TNF-α were measured using various types of immunoassays and expressed as pg/ml. Studies that had up to 5% of their samples below their assays' detection limit either imputed TNF-α serum level values with a value between 0 and the detection limit or set these values equal to the detection limit, the method of choice was left at the study's discretion. We excluded studies that had over 5% of their samples measured below the detection limit.

Genotyping and imputation

Genome wide-genotyping and subsequent quality control was performed by each of the participating studies using a variety of genotyping platforms. Next, each study performed genotype imputation using haplotypes from the Hapmap Phase II reference panel (NCBI Build 36), using IMPUTE[22], MACH[23] or Minimac[24] to infer unobserved genotypes, resulting in a per-study set of ~2.5 million variants.

Quality control

Prior to the meta-analysis, quality control of GWAS results files was carried out with the in-house built QCGWAS package[25], which performs an automated check of the data distributions, evaluates missing and invalid data, compares the alleles and allele frequencies to a reference panel, compares observed with expected p-values based on beta and SE, and creates skewness and kurtosis graphs, histograms, and QQ and Manhattan plots. Using the QCGWAS result files, cohort-specific filter thresholds for the allele frequency and imputation quality were determined if needed to normalize the inflation of statistical tests. Otherwise we did not filter for allele frequency and for imputation quality we used method-specific thresholds[26].

Statistical methods

Each cohort tested genotype associations with serum levels of TNF-α by means of linear regression under an additive model accounting for imputation uncertainty while adjusting for age, age^2, sex, body mass index (BMI), and study-specific covariates such as principal components or study site and adjusting for relatedness, when necessary. Prior to the association analyses, sample measurement values were first natural log transformed after which samples with extreme TNF-α levels (>4 S.D. from the mean) were excluded, to generate an approximately normal distribution. Results from the individual studies were pooled using an inverse variance weighted, fixed-effects meta-analysis as implemented in GWAMA[27]. We corrected for residual population substructure by applying double genomic control, i.e. first to each individual study and subsequently also to the pooled results after meta-analysis.

Based on our meta-analysis results, SNPs taken forward to the replication stage had firstly to meet the following criteria: (i) being statistically independent, (ii) having a minimum p-value ≤ 1×10^{-5} (i.e. suggestive hit), and (iii) found in at least half of the cohorts and half of the total sample size. We performed an approximate joint conditional analysis based on summary statistics implemented in GCTA[28], using high quality variants from the imputed genotype dataset of the Netherlands Study of Depression and Anxiety (NESDA)[29] study to identify statistically independent signals. Secondly, if regional association plots generated through LocusZoom[30] for our loci revealed low LD ($r^2 < 0.1$) of a statistically independent lead SNP with secondary variants in that locus, we performed pairwise LD checking for these loci in SNAP (Hapmap 22 data)[31] to verify low LD with secondary signals, and included these in the replication variant set.

Replication Study

Population

A total of 41 independent variants were taken forward for replication. Replication analyses were performed using a combination of in-silico and de novo genotyping in 7,771 individuals from European ancestry from 9 studies including the Center for Health Discovery and Wellbeing cohort (CHDWB, N=583), the Copenhagen Prospective Studies on Asthma in Childhood at-risk mother-child cohort (COPSAC2000, N=251), the Genome-Wide Population-Based Association Study of Extremely Overweight Young Adults study (GOYA, N=166), the Invecchiare in Chianti study (InCHIANTI, N=1109), the NESDA and NTR studies (additionally genotyped samples, independent samples from discovery, N=638 and N=3450 respectively), the Memory and Aging Project (MAP, N=341), the Suivi Temporaire Annuel Non Invasif de la Santé des Lorrains Assurés Sociaux study (STANISLAS, N=745) and the British Women's Heart and Health Study (BWHHS, as part of the UCLEB consortium, N=489).

Data analysis

Individual studies tested each of the selected SNPs, using the same statistical model as in the discovery association analyses. We next compared alignment of effect size estimates of all replication variants from each individual replication study against the effect size estimates from the discovery meta-analyses. When effect sizes from individual cohorts did not align, we excluded these cohorts from the replication meta-analyses. To account for differences in sensitivities and dynamic ranges of TNF-α assays used in the replication association analyses as compared to the discovery analyses, we combined results across the replication studies using a fixed-effects, sample-size weighted Z-score meta-analysis

(Stouffer's method) as implemented in the METAL package[32]. The association results from the discovery and replication were also combined using a sample-size weighted Z-score meta-analysis. Variants that were significant in the replication meta-analysis at $p<0.05$ were considered as replicated variants. Those variants that had a $p<5x10^{-8}$, with a lower p-value in the combined analysis of the discovery and replication studies as compared to their corresponding p-value in the discovery were considered genome-wide significant.

Heritability estimates

We calculated the variance explained by all independent lead SNPs from the meta-analysis using the following formula :

$$\sum_{i=1}^{n} \frac{\beta_i^2 \cdot 2 \cdot EAF_i \cdot (1 - EAF_i)}{\sigma^2(residuals(\log(TNF\alpha)))}$$

where EAF is the effect allele frequency and the effect size of the individual variants and n is the total number of lead variants. The variance of the residuals of log (TNF-α) was calculated using data from the NESDA cohort (N=2,512).

We estimated the total common SNP heritability of serum TNF-α levels explained by all GWAS variants using the observed Z-statistics from the discovery analyses for a subset of pruned SNPs within our discovery association summary statistics. Following the original method (SumVg)[33], we pruned the imputed genotype dataset of the NESDA cohort[29] using PLINK[34], removing highly correlated SNPs ($r^2>0.25$) within a 100-SNP sliding window, and a step size of 25 SNPs per move. This resulted in a pruned set of 163,459 SNPs. We used the Z-scores in the summary statistics of the discovery association for this set of variants to estimate the total explained variance for serum levels of TNF-α.

Fine mapping and identification of putative causal risk variants and candidate genes

Using 1000 Genomes sequence data (Phase1 Integrated Release, Version 3, 2012.04.30), we searched for variants in high LD of $r^2>0.8$ within a 1 Mb region on either side of the lead SNP using tools available in Liftover[35], VCFtools[36] and PLINK[34]. We subsequently annotated the remaining variants using ANNOVAR[37] with the RefSeq[38] database for variant function and genic residence or distance, and for presence in the GWAS catalog[39] to identify associations with other phenotypes.

RESULTS

A total of 23,141 individuals (56% female) of European descent from 16 studies were combined in a discovery GWAS meta-analysis with up to 2,482,219 autosomal variants passing quality control. Prior to meta-analysis, we identified genome-wide significant ($p<5x10^{-8}$) variants in the *ABO* locus in 4 individual studies, with rs644234 as the lead SNP. Each of these 4 studies, but none of the others, had used an R&D systems high-sensitivity assay kit (R&D systems, Minneapolis, MN, USA). Following earlier indications that this might be an assay specific association[20], we removed variance attributable to this locus, by conditioning the association results of these cohorts on the coded allele dosages of rs644234, after which the summary statistics were combined with those of the other 12 studies in a discovery association meta-analysis.

We found 14 variants associated with serum levels of TNF-α at genome-wide significance ($p < 5x10^{-8}$), representing three independent genetic loci on chromosome 6p21, 12q24 and 15q21. The statistically independent lead SNPs of each of these three regions are presented in Table 1, together with one LD-independent variant for our 12q24 locus. The minor allele of rs2857602 on chromosome 6p21 ($p<1.21x10^{-8}$) and rs7182229 on chromosome 15 ($p<1.15x10^{-8}$), and major allele of rs10744774 on chromosome 12q24 ($p=8.77x10^{-10}$) were positively associated with TNF-α, having per allele effect sizes from 0.030 to 0.050 increase lnTNF-α (pg/ml) levels in the discovery stage of our meta-analysis (Figure 1, Table 1). After joint conditional analysis and inspection of the LD-structure of loci in Locuszoom plots, we identified one additional LD-independent variant, rs3184504, in the 12q24 locus ($p< 2.42x10^{-08}$, $r^2=0.107$ with rs10744774). We next took all 4 statistically and LD-independent genome-wide variants forward for replication testing, plus an additional set of 37 statistically independent variants that showed suggestive association ($5\times10^{-8} <p< 1\times10^{-5}$) in the discovery analyses.

Replication analyses were performed in an additional 9 studies encompassing up to 7,771 individuals, using a combination of in-silico and de-novo genotyping and following the same QC procedures as for the discovery phase. All of the novel 4 genome-wide significant top variants in 3 loci were replicated ($p<0.05$) in the independent replication analyses (Table 1). After combining the discovery and replication analyses all independent genome-wide variants identified in the discovery increased in significance. In both

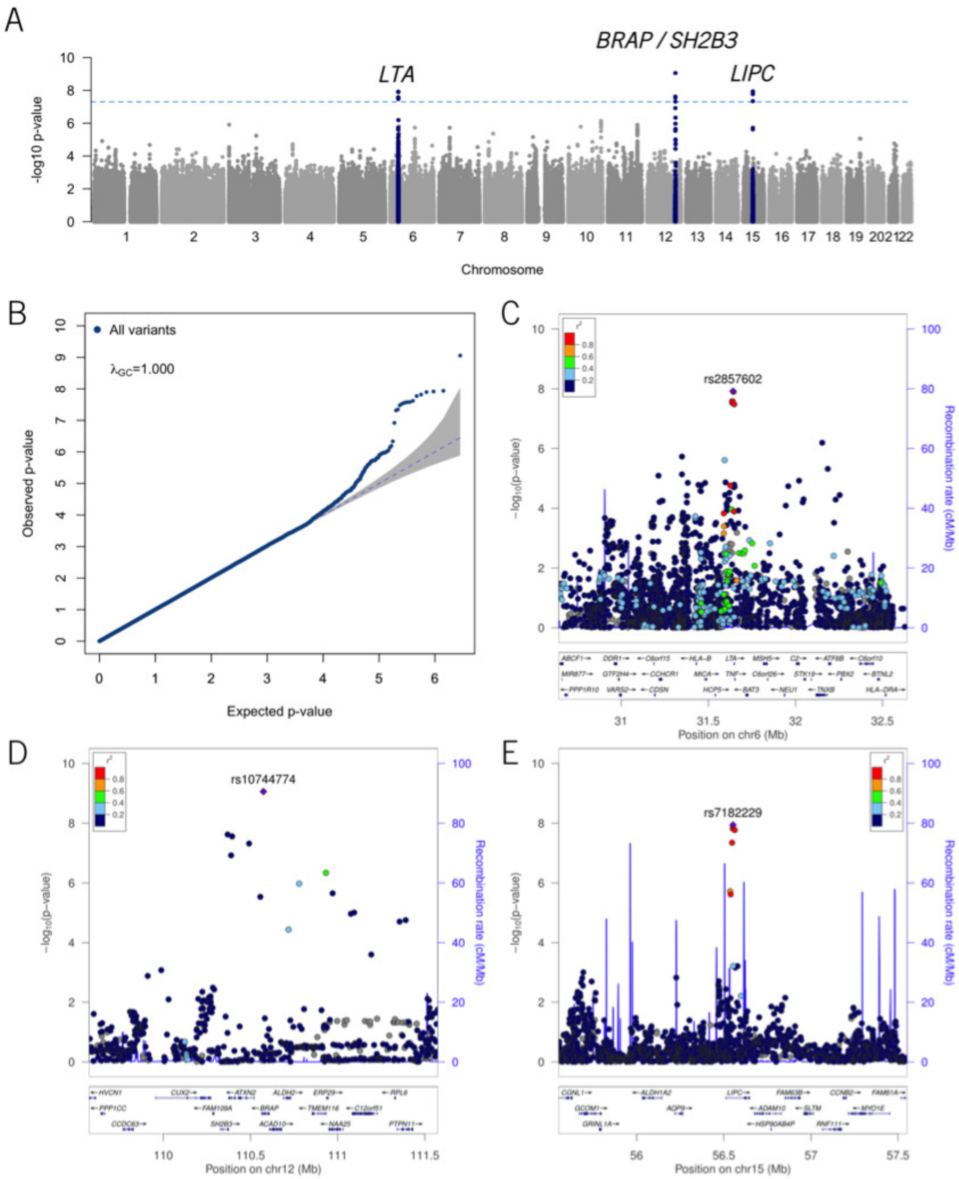


Figure 1. Manhattan plot and Locuszoom plots of the discovery analyses. Manhattan plot showing the association of SNPs with TNF-α. Loci coloured in red or blue, 3 in total, represent those for which the lead SNPs reach genome-wide significance ($P=5\times10^{-8}$) or lower. Horizontal axis : relative genomic position of variants on the genome, vertical axis : -log10 p-value of each SNP; b) Quantile-quantile plot for p-values obtained from meta-analysis. The horizontal and vertical axes represents the expected distribution of -log10(P-values) under the null hypothesis of no association, whereas the vertical axis shows the observed -log10(P-values). The blue dashed line represents the null, and λ_{gc} value represents the genomic inflation factor lambda. Each data point represents the observed versus the expected p-value of a variant included in the association analyses; c-e) Regional association plots for each of the 3 genome-wide significant loci, 6p21, 12q24 and 15q21 respectively. Pairwise LD (r^2) with the lead SNP is indicated following a color-coded scale. Horizontal axis : relative genomic position of variants within the locus, vertical axis : -log10 p-value of each SNP.

discovery and replication association analyses effect sizes were generally consistent across individual studies for genome-wide significant variants, and we did not observe evidence of heterogeneity, except only moderately for one of our lead SNPs, rs2857602, having an I^2 value of 50.2%, and Q value of 0.014. None of the 37 independent variants that showed suggestive association in the discovery stage reached genome-wide significance in the combined discovery and replication analyses.

Using the SumVg method[33], we estimate the total variance explained by all GWAS variants for TNF-α to be 3.83%. Our 4 lead SNPs combined explain approximately 0.58% of the genetic variance of levels of TNF-α using data from the NESDA cohort.

Table 1. 4 independent variants in 3 genomic loci associated with serum-levels of TNF-α.

Locus	Variant	E/N	EAF	β(SE)	$P_{discovery}$	$P_{replication}$	$P_{combined}$	Function	Genes
6p21	rs2857602	G/A	0.38	0.032 (0.006)	$1.21x10^{-8}$	$6.31x10^{-5}$	$3.30x10^{-12}$	intergenic	*LTA*
12q24	rs3184504	T/C	0.48	0.030 (0.005)	$2.42x10^{-8}$	$5.33x10^{-3}$	$3.96x10^{-10}$	missense	*SH2B3*
	rs10744774	A/C	0.83	0.044 (0.007)	$8.77x10^{-10}$	$2.93x10^{-2}$	$6.94x10^{-11}$	intronic	*BRAP*
15q21	rs7182229	T/G	0.11	0.050 (0.009)	$1.15x10^{-8}$	$2.52x10^{-2}$	$1.07x10^{-9}$	intronic	*LIPC*

Variants are shown that reached $P < 5×10^{-8}$ in the combined analysis and are independent lead SNPs. Sample sizes: discovery cohorts, n=23,141; replication cohorts, n=7,771; combined, n=30,912. The effect sizes (b) in the discovery phase, given for the effect allele. E/N; E is the effect allele, and N is the non-effect allele. Effect sizes and standard error (SE) values are in natural log (pg/ml) units.

Fine mapping and identification of putative causal risk variants

Our 4 lead variants in 3 loci were located within the vicinity of Tumor necrosis factor – Lymphotoxin alpha (*TNF-LTA*, rs2857602, 6p21, 6498bp downstream of *LTA*), SH2B adaptor protein 3 (*SH2B3*, rs3184504, 12q24, exonic), BRCA1 associated protein (*BRAP*, rs10744774, 12q24, intronic) and hepatic lipase (*LIPC*, rs7182229, 15q21, intronic). The variant in *LIPC*, rs7182229, also resides in an overlapping non-coding RNA gene (*LOC101928694*). Our lead SNP in the 6p21 resides in the promoter region of *LTA* is in high LD (r^2>0.8) with variants that reside within intronic regions of Nuclear Factor of Kappa light polypeptide gene enhancer in B-cells Inhibitor-Like 1 (*NFKBIL1*) and intronic regions or the 5'UTR of *LTA*. Among the four lead SNPs and associated (r^2>0.8) look-up variants, only the chromosome 12 rs3184504 variant itself, was a non-synonymous SNP, whereas high LD SNPs map to intronic SNPs in *BRAP* and *ATNX2*. The lead variant on the chromosome 15 locus and its variants in high LD are all intronic and map to *LIPC* and *LOC101928694*.

A lookup of the 4 lead variants and SNPs in high LD in the GWAS catalog (accessed October 12, 2015) revealed that only the non-synonymous lead SNP rs3184504 in the chromosome 12 locus and a large proportion of high-LD variants were previously found to be associated with a range of other traits and diseases, including common cardiovascular traits and diseases (such as red blood cell traits, coronary heart disease, platelet counts, diastolic and systolic blood pressure) and auto-immune diseases (such as rheumatoid arthritis, type 1 diabetes and celiac disease).

DISCUSSION

This is the first meta-GWAS analysis of serum levels of TNF-α in a total of 30,912 individuals of European descent. We identified 4 SNPs in 3 loci implicating (nearby) genes that are associated with serum levels of TNF-α and that may hold clues to their involvement in a plethora of diseases known to be related to TNF-α levels.

At the chromosome six locus, the TNF-α associated rs2857602 SNP resides in the promoter region of *LTA*. *LTA* is a most relevant candidate gene for a number of reasons. Firstly, a few candidate gene studies have found polymorphisms residing within *LTA* affect serum levels of TNF-α [40,41], while other studies linked the promoter region to inflammatory diseases [42,43]. Secondly, *TNF* and *LTA* are genes that have comparable biological activities, have 35% identity and 50% homology in the amino acid sequence, and also have receptors in common, i.e., the TNF-α receptors 1 and 2, on a range of leukocytes and parenchymal cells[44,45]. Both are key mediators in the immune response and synthesized by activated monocytes and lymphocytes. In addition, it has been shown previously that a variant in the *TNF* promoter region, in literature often referred to as *TNF*:-308G/A (dbSNP ID: rs1800629), influences the binding of RNA polymerase II to and allele-specific transcription of *LTA*[46].

Several genes at the 12q24 locus make plausible biological candidates. This region maps to one of the largest blocks of LD in the human genome, spanning over 1Mb and harbouring several genes[47]. Within this block, rs3184504 is an exonic variant in

SH2B3, which encodes the SH2B adapter protein 3, a negative regulator of cytokine signalling[48] and *TNF* signalling in particular for endothelial cells[49]. This variant has also been found to be significantly associated with a diverse set of human complex traits, including multiple autoimmune disorders[50–54] and cardiovascular traits[45–50]. However, two common pathogenicity prediction tools predict this variant to be benign (PolyPhen2[55]) and tolerated (SIFT[56]). Interestingly, we also identified a variant, rs10744774, that is independent from rs3184504 in this block, and that maps to *BRAP*, which is a cytoplasmic protein that regulates nuclear targeting by retaining proteins with a nuclear localization signal in the cytoplasm[57] and is involved in MAPK signalling[58]. Not much is known about the association between *BRAP* and *TNF*, other than that *BRAP* is involved in the nuclear translocation of NF-kB following TNF-α stimulation[59]. NF-kB is a transcription factor that once activated, translocates from the cytoplasm to the nucleus and binds to promoter/ enhancer regions of a wide variety of genes including those that regulate a range of inflammatory and immune responses.

The rs7182229 variant in the 15q21 locus is an intronic variant in *LIPC*, a gene that encodes for hepatic lipase which is an enzyme synthesized and secreted by the liver that catalyses the hydrolysis of triglycerides and phospholipids, but is also involved in receptor-mediated lipoprotein uptake into the liver. HDL regulates the release of hepatic lipase from the liver and HDL structure controls HDL transport and activation in the circulation. By now it is well known that TNF-α regulates and interferes with lipid homeostasis[60,61], more specifically with triglyceride and cholesterol metabolism[62,63].

Several large GWAS meta-analyses have identified variation in *LIPC*, associated with a wide variety of biomarkers and disease, including lipid traits[64–66], haematological traits[67,68], metabolic syndrome[69] and age-related macular degeneration[70]. Nevertheless, none of the lead SNPS for these traits is in high with our lead *LIPC* variant ($r^2 < 0.1$, data not shown). At the same time, our lead variant maps to an overlapping ncRNA gene, *LOC101928694*, of currently unknown function. Our search for additional coding variants in 1000 Genomes data with high LD ($r^2 > 0.8$) with lead SNPs at our loci revealed no additional variants with potential effect on protein function. Instead, some of our lead variants may influence TNF-α levels through other regulatory mechanisms. In particular for the *TNF-LTA* locus it is well established by now that post-translational histone modifications, DNA methylation states, and higher-order chromatin interactions control regulation of cytokines genes within the MHC region[71]. *TNF* itself is known to induce surface expression of class II MHC molecules in a wide variety of cell types[72–76], and also has shown to modulate MHC I class antigen processing and presentation[77]. The MHC region remains one of the most complex regions in the genome to study, owing to its large linkage disequilibrium blocks, sequence diversity and high density of genes[78]. Therefore, pinpointing causal variants associated with serum levels of TNF-α within this region remains challenging.

Our study has a few limitations. First, though we have achieved substantially higher sample sizes as compared to previous studies, we were able to only explain less than half a percent of the total heritability. A very conservative estimate for total SNP-heritability amounts to 3.8% in our study, as total heritability estimates from twin studies range up to 39%, therefore, leave much room for additional genetic variation influencing TNF-α levels to be discovered, requiring sizable increases in sample sizes as compared to our study. Secondly, we have included studies that employed a variety of different assays to measure TNF-α levels. Even though our identified variants showed consistency of effect size and direction, there were indications of assay-specific effects. Ideally, efforts should be made to harmonize TNF-α level measurements. Prior to analyses, assay performance should be assessed, and ideally one assay type and manufacturer should be chosen, though this will be prove to be difficult to realize in practice. Unlike relatively more straightforward phenotype measurements such as anthropomorphic traits, reliable measurements of cytokines such as TNF-α depend on many factors, such as the quality of the antibody used, how antibodies were generated (peptide or whole molecule) and how blood samples were treated. Harmonized protocols to measure TNF-α levels therefore would be of great value, and reducing the variability in levels measured due to non-genetic factors, reducing confounding and increasing power. Thirdly, though the genetic variation analysed in this study comprises a substantial set of 2,5 million variants based on Hapmap imputation, more recent sequence efforts, such as those of the Haplotype Reference Consortium[79], the 100,000 Genomes Project[80] and, the Sequencing IsoLates Consortium (SILC) for isolated populations will allow imputations on a much richer and denser set of genetic variants, allowing he discovery of novel genetic variants and improved fine-mapping to identify variants with more likely functional impact and causal involvement.

In summary, these results substantially extend the knowledge on genetic determination of serum TNF-α levels. We present the first set of variants genome-wide significantly associated with TNF-α, and replicating in independent samples. At the variant level, determination of TNF-α serum levels appears to be affected by a variety of mechanisms ; not only by changes in protein coding, but also by variants that may alter regulatory elements within or near our loci. Furthermore, a diversity of genes with differing functions is implicated in our study. Genes harboured by our loci, and in certain cases also our lead SNPs or high LD variants appear to be pleiotropic in nature, implicated in intermediate traits, such as lipid and red blood cell traits, but also in disease endpoints, not only further confirming a role of TNF-α a as a biomarker for many diseases, but in many cases also shared genetic components in previously unsuspected genes.

Web resources

QCGWAS, https://cran.r-project.org/web/packages/QCGWAS/index.html
GWAMA, http://www.well.ox.ac.uk/gwama/
METAL, http://csg.sph.umich.edu//abecasis/metal/
GCTA, http://www.complextraitgenomics.com/software/gcta/
LocusZoom, http://csg.sph.umich.edu/locuszoom/
1000 Genomes, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
VCFtools, http://vcftools.sourceforge.net/
ANNOVAR, http://www.openbioinformatics.org/annovar/
Polyphen2, http://genetics.bwh.harvard.edu/pph2/
SIFT, http://sift.jcvi.org/

REFERENCES

1. Cawthorn, W. P. & Sethi, J. K. TNF-alpha and adipocyte biology. FEBS Lett. 582, 117–131 (2008).
2. Kriegler, M., Perez, C., DeFay, K., Albert, I. & Lu, S. D. A novel form of *TNF*/cachectin is a cell surface cytotoxic transmembrane protein: ramifications for the complex physiology of *TNF*. Cell 53, 45–53 (1988).
3. Black, R. A. et al. A metalloproteinase disintegrin that releases tumour-necrosis factor-alpha from cells. Nature 385, 729–733 (1997).
4. Tracey, D., Klareskog, L., Sasso, E. H., Salfeld, J. G. & Tak, P. P. Tumor necrosis factor antagonist mechanisms of action: a comprehensive review. Pharmacol. Ther. 117, 244–279 (2008).
5. Moelants, E. A. V., Mortier, A., Van Damme, J. & Proost, P. Regulation of *TNF-* with a focus on rheumatoid arthritis. Immunol. Cell Biol. 91, 393–401 (2013).
6. Targownik, L. E. & Bernstein, C. N. Infectious and malignant complications of *TNF* inhibitor therapy in IBD. Am. J. Gastroenterol. 108, 1835–1842, quiz 1843 (2013).
7. García-Trejo, A. R. et al. Tumor necrosis factor alpha promoter polymorphisms in Mexican patients with dengue fever. Acta Trop. 120, 67–71 (2011).
8. Ridker, P. M. et al. Elevation of tumor necrosis factor-alpha and increased risk of recurrent coronary events after myocardial infarction. Circulation 101, 2149–2153 (2000).
9. Koller-Strametz, J. et al. Circulating tumor necrosis factor-alpha levels in chronic heart failure: relation to its soluble receptor II, interleukin-6, and neurohumoral variables. J. Heart Lung Transplant. 17, 356–362 (1998).
10. Swardfager, W. et al. A meta-analysis of cytokines in Alzheimer's disease. Biol. Psychiatry 68, 930–941 (2010).
11. Dowlati, Y. et al. A meta-analysis of cytokines in major depression. Biol. Psychiatry 67, 446–457 (2010).
12. Antoni, C. & Braun, J. Side effects of anti-*TNF* therapy: current knowledge. Clin. Exp. Rheumatol. 20, S152–157 (2002).
13. van Schouwenburg, P. A., Rispens, T. & Wolbink, G. J. Immunogenicity of anti-*TNF* biologic therapies for rheumatoid arthritis. Nat Rev Rheumatol 9, 164–172 (2013).
14. Sas, A. A. et al. The age-dependency of genetic and environmental influences on serum cytokine levels: a twin study. Cytokine 60, 108–113 (2012).
15. Neijts, M. et al. Genetic architecture of the pro-inflammatory state in an extended twin-family design. Twin Res Hum Genet 16, 931–940 (2013).
16. Elahi, M. M., Asotra, K., Matata, B. M. & Mastana, S. S. Tumor necrosis factor alpha -308 gene locus promoter polymorphism: an analysis of association with health and disease. Biochim. Biophys. Acta 1792, 163–172 (2009).
17. Mekinian, A. et al. Functional study of *TNF-* promoter polymorphisms: literature review and meta-analysis. Eur. Cytokine Netw. 22, 88–102 (2011).
18. Stüber, F., Petersen, M., Bokelmann, F. & Schade, U. A genomic polymorphism within the tumor necrosis factor locus influences plasma tumor necrosis factor-alpha concentrations and outcome of patients with severe sepsis. Crit. Care Med. 24, 381–384 (1996).
19. Whichelow, C. E., Hitman, G. A., Raafat, I., Bottazzo, G. F. & Sachs, J. A. The effect of

TNF*B gene polymorphism on *TNF*-alpha and -beta secretion levels in patients with insulin-dependent diabetes mellitus and healthy controls. Eur. J. Immunogenet. 23, 425–435 (1996).

20. Melzer, D. et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 4, e1000072 (2008).

21. Wood, A. R. et al. Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. PLoS ONE 8, e64343 (2013).

22. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913 (2007).

23. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34, 816–834 (2010).

24. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).

25. van der Most, P. J. et al. QCGWAS: A flexible R package for automated quality control of genome-wide association results. Bioinformatics (2014). doi:10.1093/bioinformatics/btt745

26. Pei, Y.-F., Zhang, L., Li, J. & Deng, H.-W. Analyses and Comparison of Imputation-Based Association Methods. PLoS ONE 5, e10827 (2010).

27. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11, 288 (2010).

28. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569 (2010).

29. Penninx, B. W. J. H. et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatr Res 17, 121–140 (2008).

30. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337 (2010).

31. Johnson, A. D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24, 2938–2939 (2008).

32. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).

33. So, H.-C., Li, M. & Sham, P. C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. Genet. Epidemiol. 35, 447–456 (2011).

34. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).

35. Hinrichs, A. S. et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34, D590–598 (2006).

36. Danecek, P. et al. The variant call format and VCFtools. Bioinformatics 27, 2156–2158 (2011).

37. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).

38. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a

curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33, D501–504 (2005).

39. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42, D1001–1006 (2014).

40. Tomasdottir, H. et al. Tumor necrosis factor gene polymorphism is associated with enhanced systemic inflammatory response and increased cardiopulmonary morbidity after cardiac surgery. Anesth. Analg. 97, 944–949, table of contents (2003).

41. Warzocha, K. et al. Genetic Polymorphisms in the Tumor Necrosis Factor Locus Influence Non-Hodgkin's Lymphoma Outcome. Blood 91, 3574–3581 (1998).

42. Lee, K.-A. et al. Association between a polymorphism in the lymphotoxin-a promoter region and migraine. Headache 47, 1056–1062 (2007).

43. Ali, S. et al. Association of variants in BAT1-*LTA-TNF*-BTNL2 genes within 6p21.3 region show graded risk to leprosy in unrelated cohorts of Indian population. Hum. Genet. 131, 703–716 (2012).

44. Aggarwal, B. B., Eessalu, T. E. & Hass, P. E. Characterization of receptors for human tumour necrosis factor and their regulation by gamma-interferon. Nature 318, 665–667 (1985).

45. Hehlgans, T. & Pfeffer, K. The intriguing biology of the tumour necrosis factor/tumour necrosis factor receptor superfamily: players, rules and the games. Immunology 115, 1–20 (2005).

46. Knight, J. C., Keating, B. J., Rockett, K. A. & Kwiatkowski, D. P. In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. Nat. Genet. 33, 469–475 (2003).

47. Scherer, S. E. et al. The finished DNA sequence of human chromosome 12. Nature 440, 346–351 (2006).

48. Velazquez, L. et al. Cytokine signaling and hematopoietic homeostasis are disrupted in Lnk-deficient mice. J. Exp. Med. 195, 1599–1611 (2002).

49. Fitau, J., Boulday, G., Coulon, F., Quillard, T. & Charreau, B. The adaptor molecule Lnk negatively regulates tumor necrosis factor-alpha-dependent VCAM-1 expression in endothelial cells through inhibition of the ERK1 and -2 pathways. J. Biol. Chem. 281, 20148–20159 (2006).

50. Barrett, J. C. et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. 41, 703–707 (2009).

51. Plagnol, V. et al. Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLoS Genet. 7, e1002216 (2011).

52. Alcina, A. et al. The autoimmune disease-associated KIF5A, CD226 and *SH2B3* gene variants confer susceptibility for multiple sclerosis. Genes Immun. 11, 439–445 (2010).

53. Stahl, E. A. et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42, 508–514 (2010).

54. Izzo, V. et al. Improving the estimation of celiac disease sibling risk by non-HLA genes. PLoS ONE 6, e26920 (2011).

55. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).

56. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT

algorithm. Nat Protoc 4, 1073–1081 (2009).

57. Li, S. et al. Identification of a novel cytoplasmic protein that specifically binds to nuclear localization signal motifs. J. Biol. Chem. 273, 6183–6189 (1998).

58. Lanctot, A. A., Peng, C.-Y., Pawlisz, A. S., Joksimovic, M. & Feng, Y. Spatially dependent dynamic MAPK modulation by the Nde1-Lis1-*BRAP* complex patterns mammalian CNS. Dev. Cell 25, 241–255 (2013).

59. Takashima, O. et al. *BRAP*2 regulates temporal control of NF- B localization mediated by inflammatory response. PLoS ONE 8, e58911 (2013).

60. Fon Tacer, K., Kuzman, D., Seliskar, M., Pompon, D. & Rozman, D. TNF-alpha interferes with lipid homeostasis and activates acute and proatherogenic processes. Physiol. Genomics 31, 216–227 (2007).

61. Chen, X., Xun, K., Chen, L. & Wang, Y. TNF-alpha, a potent lipid metabolism regulator. Cell Biochem. Funct. 27, 407–416 (2009).

62. Grunfeld, C. & Feingold, K. R. Tumor necrosis factor, cytokines, and the hyperlipidemia of infection. Trends Endocrinol. Metab. 2, 213–219 (1991).

63. Memon, R. A., Grunfeld, C., Moser, A. H. & Feingold, K. R. Tumor necrosis factor mediates the effects of endotoxin on cholesterol and triglyceride metabolism in mice. Endocrinology 132, 2246–2253 (1993).

64. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010).

65. Global Lipids Genetics Consortium et al. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274–1283 (2013).

66. Demirkan, A. et al. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. PLoS Genet. 8, e1002490 (2012).

67. Kamatani, Y. et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat. Genet. 42, 210–215 (2010).

68. van der Harst, P. et al. Seventy-five genetic loci influencing the human red blood cell. Nature 492, 369–375 (2012).

69. Kristiansson, K. et al. Genome-wide screen for metabolic syndrome susceptibility Loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits. Circ Cardiovasc Genet 5, 242–249 (2012).

70. Fritsche, L. G. et al. Seven new loci associated with age-related macular degeneration. Nat. Genet. 45, 433–439, 439e1–2 (2013).

71. Taylor, J. M., Wicks, K., Vandiedonck, C. & Knight, J. C. Chromatin profiling across the human tumour necrosis factor gene locus reveals a complex, cell type-specific landscape with novel regulatory elements. Nucleic Acids Res. 36, 4845–4862 (2008).

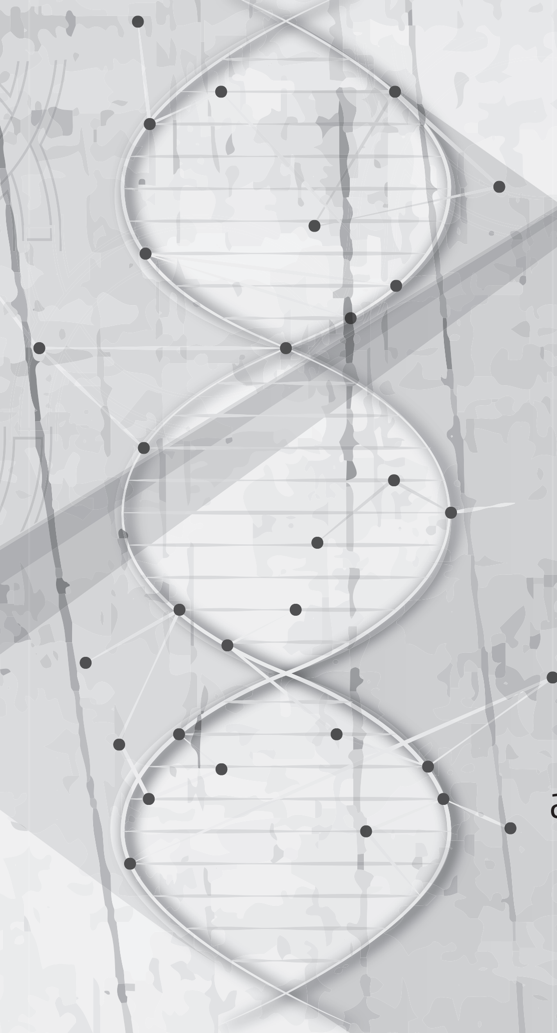
72. Chang, R. J. & Lee, S. H. Effects of interferon-gamma and tumor necrosis factor-alpha on the expression of an Ia antigen on a murine macrophage cell line. J. Immunol. 137, 2853–2856 (1986).

73. Collins, T., Lapierre, L. A., Fiers, W., Strominger, J. L. & Pober, J. S. Recombinant human tumor necrosis factor increases mRNA levels and surface expression of HLA-A,B antigens in vascular endothelial cells and dermal fibroblasts in vitro. Proc Natl Acad Sci U S A 83, 446–450 (1986).

74. Marth, C., Zech, J., Böck, G., Mayer, I. & Daxenbichler, G. Effects of retinoids and

interferon-gamma on cultured breast cancer cells in comparison with tumor necrosis factor alpha. Int. J. Cancer 40, 840–845 (1987).

75. Massa, P. T., Schimpl, A., Wecker, E. & ter Meulen, V. Tumor necrosis factor amplifies measles virus-mediated Ia induction on astrocytes. Proc. Natl. Acad. Sci. U.S.A. 84, 7242–7245 (1987).
76. Pujol-Borrell, R. et al. HLA class II induction in human islet cells by interferon-gamma plus tumour necrosis factor or lymphotoxin. Nature 326, 304–306 (1987).
77. Hallermalm, K. et al. Tumor necrosis factor-alpha induces coordinated changes in major histocompatibility class I presentation pathway, resulting in increased stability of class I complexes at the cell surface. Blood 98, 1108–1115 (2001).
78. de Bakker, P. I. W. & Raychaudhuri, S. Interrogating the major histocompatibility complex with high-throughput genomics. Hum. Mol. Genet. 21, R29–36 (2012).
79. The Haplotype Reference Consortium. at <http://www.haplotype-reference-consortium.org>
80. Genomics England is delivering the 100,000 Genomes Project. at http://www.genomicsengland.co.uk.



SECTION I :

GWAS AND INFLAMMATORY MARKER GENETICS

4

A genome-wide association scan identifies novel loci associated with Interleukin-6 levels

On behalf of the IL-6 Meta-GWAS Consortium:

Bram P. Prins, co-authors of the IL-6 Meta-GWAS Consortium, co-authors of the CHARGE Inflammation Working Group, Harold Snieder and Behrooz Z. Alizadeh

ABSTRACT

Interleukin-6 (IL-6) is a cytokine with both pro- and anti-inflammatory properties, is highly multifunctional in nature, and is synthesized by a wide range of different tissues and cell types. Increased levels of IL6 have been associated with various classes of diseases and plays major roles in the aetiologies of autoimmune and cardiovascular diseases. Even though IL-6 levels are highly heritable with estimates up to 61%, only a few genetic loci involved in IL-6 levels in blood have been identified. We conducted a meta-analysis of genome-wide association studies of serum concentration of IL-6 encompassing 66,341 individuals of European descent and identified three independent genome-wide significantly associated loci at chromosome 1q21, 2p14, and 6p21. Our loci harbour well-known inflammation related genes, including *IL6R* (1q21,rs4537545, p=1.20x10^{-122}), *IL1F10* and *IL1RN* (2q14, rs6734238, p=1.84x10^{-11}), and *HLA-DRB1/DRB5* (6p21, rs660895, p=1.56x10^{-10}). These genes have previously been associated with a variety of diseases, including asthma, rheumatoid arthritis, and IgA nephropathy and serum levels of other biomarkers like C-Reactive Protein (CRP), fibrinogen, and soluble Interleukin-6 receptor (sIL-6R). Our study increases the number of loci that are known to be genome-wide significantly associated with IL-6 from 1 to 3, and provides additional insights in the biology, such as the involvement of the HLA complex, and ample overlap of our loci with various traits and disease classes.

INTRODUCTION

Interleukin-6 (IL-6) is a highly multifunctional cytokine that is a member of the interleukin cytokine family, which is involved in a wide range of cellular processes, from migration and adhesion to proliferation and maturation[1,2]. Specifically regarding the immune system interleukin cytokines have modulatory roles and are involved in cell differentiation and activation. IL-6 itself is synthesized by a wide variety of different cell types, both those belonging to the immune system such as monocytes[3], B-cells[4] and T-cells[5], and following stimulation of other non-immune system cells such as epithelial and smooth muscle cells[6], adipocytes[7], endothelial cells[8], and even osteoblasts[9]. In its first stages, IL-6 is synthesized in the form of a precursor protein of 212 amino acids in length, whereas the mature protein is 185 amino acids in length after proteolytic cleaving. Due to post-translational modifications such as phosphorylation and glycosylation, the final molecular form of the protein has molecular masses between 21.5-28 kilodalton.

IL-6 signalling occurs through forming a complex with two other molecules, namely the membrane-bound IL-6 receptor (mIL-6R), and gp130. Upon binding to mIL-6R, IL-6 induces homodimerization of gp130. The resulting receptor complex of IL-6, gp130, and mIL-6R activates Janus Kinases (JAKs), which in turn phosphorylate the cytoplasmic domain of gp130. This activated complex then in turn is able to trigger several pathways through which IL-6 exerts its biological activities, namely the JAK/STAT pathway, Ras/Raf pathway, and the PI3K/AKT pathway. This process is referred to as classic IL-6 signalling[10].

Increased levels of IL-6 have been observed in various disease classes, not surprisingly in autoimmune diseases such as rheumatoid arthritis (RA)[11] and systemic juvenile idiopathic arthritis (SJIA)[12], but also in cardiovascular disorders such as congestive heart failure (CHF) and coronary heart disease (CHD)[13], and even in psychological disorders such as major depressive disorder (MDD). Given its critical role in in the pathogenesis of different disorders, it forms an obvious choice for drug targeting[14], the most well known IL-6 inhibitor being tocilizumab[15]. This monoclonal antibody binds to the IL-6 receptor, which subsequently blocks IL-6 signalling by preventing the dimerization of the IL-6/IL-6R complex with membrane-bound gp130. It has been shown to have high efficacies in patients with RA[16] and SJIA[17], but it is not used for non-immune disorders. As with most immunosuppressive agents, one of the more serious side effects includes increased infection rates[18] (in particular in the upper respiratory tract), whilst at the same time masking certain usual symptoms of infection (e.g. high temperatures).

Baseline levels of IL-6 are highly heritable (heritability estimates from twin studies ranging from 0.15 to 0.61[19–21]. To date a handful of relatively small-scale genome-wide association studies have been performed for IL-6[22–25], and even a whole genome sequence based association study[26], identifying variants reaching genome-wide significance in the IL-6 receptor gene (*IL-6R*) and the gene encoding Histo-blood group ABO system transferase (*ABO*). A genetic risk score constructed of variants identified in the study by Shah and colleagues[25] explained up to 2% of variation in IL-6 levels[25], leaving a substantial part of the heritability unexplained.

In this study we try to overcome power limitations of previous GWASs by substantially increasing the sample size through inclusion of more cohorts and identify additional genetic variation explaining IL-6 levels. We present the results of a large-scale meta-analysis of GWASs with a combined sample size of 66,341 samples, followed by fine-mapping and bioinformatic analyses of genetic loci and variants.

METHODS

Discovery study

Study population

The discovery stage included 52,654 individuals from 26 cohorts of European ancestry (Supplementary Table 1): the Avon Longitudinal Study of Parents and Children (ALSPAC, N=4129), the Amish study (Amish, N=489), the Atherosclerosis Risk in Communities study (ARIC, N=511), the Baltimore Longitudinal Study of Aging (BLSA, N=477), the Cardiovascular Health Study (CHS, N=3028), Cohorte Lausannoise (CoLaus, 4,938), the Family Heart Study (FamHS, N=1,304), the Framingham Heart Study (FHS, N=6,858), Genetics of Lipid Lowering Drugs and Diet Network (GOLDN, N=821), Health ABC (HABC, N=1,597), the Helsinki Birth Cohort Study (HBCS, N=1,716), the Invecchiare in Chianti study (InCHIANTI, N=1,208) , the Lothian Birth Cohort of 1921 (LBC1921, N=166), the Lothian Birth Cohort of 1936 (LBC1936, N=759), the Leiden Longevity Study (LLS, N=1798), the Ludwigshafen Risk and Cardiovascular Health study (LURIC, N=604), the Monitoring of Trends and Determinants in Cardiovascular Disease / Kooperative Gesundheitsforschung in der Region Augsburg studies (MONICA/KORA, N=1,625), Mister Osteoporosis Sweden study (MrOS Swe, N=938), the Netherlands Study of Depression and Anxiety (NESDA, N=1,746), the Netherlands Twin Registry (NTR, N=3,152), the The PHArmacogenetic study of Statins in the Elderly at risk/PROspective Study of Pravastatin in the Elderly at Risk study (PROSPER/PHASE, N=5,130), the Rotterdam Study (RS, N=599), the National Institute on Aging (NIA) SardiNIA Study (SardiNIA, N=4,621), the Study of Health in Pomerania study (SHIP, N=1,327), TwinsUK (N=1,103) and the Young Finns Study (YFS, N=2,017). Only population-based samples or healthy controls from case-control studies were included in the analyses.

IL-6 serum level measurements

Each study typically collected venous blood samples from their participants frozen as either serum or plasma and stored below -80°C until the time of measurement. Serum or plasma levels of IL-6 were measured using various types of immunoassays and expressed as pg/ml. For studies with up to 5% of their samples below the assay' s detection limit, IL-6 serum level values were either imputed with a random value between 0 and the detection limit or they were set equal to the limit of detection (LOD), the choice of method was left at the analysts' discretion. Cohorts with over 5% of their samples measured below the LOD were asked to perform a survival-based association analysis using a method proposed by Dinse et. al[27], as substituting non-detects (NDs) with a generated value smaller than or equal to the lower LOD insufficiently accounts for the information provided by NDs[28]. The method was implemented in the lodGWAS package[29]. Cohorts that had over 10% NDs were excluded from the analyses.

Genotyping and imputation

Genome-wide genotyping and subsequent quality control was performed by each of the participating studies using a variety of genotyping platforms. Next, each study performed genotype imputation using haplotypes from the Hapmap Phase II reference panel (NCBI Build 36), using IMPUTE[30], MACH[31], Minimac[32], or BIMBAM[33], to infer unobserved genotypes, resulting in a per-study set of ~2.5 million variants.

Statistical methods

Each cohort tested genotype associations with serum levels of IL-6 by means of linear regression under an additive SNP model accounting for imputation uncertainty while adjusting for age and sex, and study-specific covariates such as principal components or study site and for relatedness, when necessary. Prior to the association analyses, sample measurement values were first natural log transformed after which samples with extreme IL-6 levels (> 4 S.D. from the mean) were excluded, in order to generate an approximately normal distribution.

Prior to the meta-analysis, quality control of GWAS results files was carried out with the QCGWAS package in R[34], which performs an automated check of the parameter distributions, evaluates missing and invalid data, compares the alleles and allele frequencies to a reference panel (HapMap Phase II), compares observed with expected p-values based on beta and SE, and creates skewness and kurtosis graphs, precision plots, histograms, and QQ and Manhattan plots. Using the QCGWAS result files, cohort-specific filter thresholds for the allele frequency and imputation quality were determined if needed to normalize the inflation of statistical tests. Otherwise we did not filter for allele frequency and for imputation quality we used method-specific thresholds[35].

Being aware of the potential false-positive association in the *ABO* region on chromosome 9[21,22], seen for other cytokine GWAS only when using an R&D systems high-sensitivity assay kit (R&D systems, Minneapolis, MN, USA), we inspected the GWAS results from individual cohorts that had used this particular kit. When we identified genome-wide significant results in this region for a certain study, we asked analysts to condition their analyses on the top SNP in this locus.

Results from the individual studies were pooled using an inverse variance weighted, fixed-effects meta-analysis as implemented in GWAMA[36]. We corrected for residual population substructure by applying double genomic control, i.e. first to each study individually and subsequently also to the pooled results after meta-analysis.

Based on our meta-analysis results, SNPs taken forward to the replication stage had firstly to meet the following criteria: (i) being statistically independent, (ii) having a minimum p-value ≤ 1 x 10-5 (i.e. suggestive hit), and (iii) found in at least half of the cohorts and half of the total sample size. We performed an approximate joint conditional analysis based on summary statistics implemented in GCTA[37],using high quality variants from the imputed genotype dataset of the Netherlands Study of Depression and Anxiety (NESDA)[38] study to identify statistically independent signals.

Replication study

Study population

A total of 12 independent SNPs were taken forward for replication. Replication analyses were performed using a combination of in-silico and de novo genotyping in 14,774 individuals from European ancestry from 12 cohorts, including the Genomics of Overweight Young Adults study (GOYA, N=318), the Hypercoagulability and Impaired Fibrinolytic function MECHanisms predisposing to myocardial infarction (MI) study (HIFMECH, N=413), the Sydney Memory and Ageing Study (SMAS, N=847), the Netherlands Twin Register (NTR, N=3,322), the Older Australian Twin Study (OATS, N=376) , the Queensland Institute of Medical Research Asthma & Allergy Study (QIMR, N=325), the Suivi Temporaire Annuel Non Invasif de la Santé des Lorrains Assurés Sociaux study (STANISLAS, N=744) , and cohorts as part of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol Consortium (UCLEB), which are : the British Regional Heart Study (BRHS,N=2,301), the British Women's Heart and Health Study (BWHHS, N=1,904), the Caerphilly Prospective Study (CAPS, N=734), the Edinburgh Artery Study (EAS, N=540), and the Whitehall II study (WHII, N=2,959).

Statistical methods

Individual studies tested each of the selected SNPs, using the same statistical model as for the discovery association analyses. Effect size estimates of all replication variants from each individual replication study were compared with the effect size estimates from the discovery meta-analyses. When effect sizes from individual cohorts did not align, we excluded these cohorts from the replication meta-analyses (3 in total). To account for differences in sensitivities and dynamic ranges of IL-6 assays used in the replication association analyses as compared to the discovery analyses, we combined results across the replication studies using a fixed-effects, sample-size weighted Z-score meta-analysis as implemented in the METAL package[39]. The association results from the discovery and replication were also combined using a sample-size weighted Z-score meta-analysis. Variants that were significant in the replication meta-analysis ($p<0.05$) and that had a $p<5x10^{-8}$ in the combined analysis of the discovery and replication studies with a lower p-value than in the discovery meta-analysis were considered genome-wide significant.

Heritability estimates

We calculated the variance explained by all independent lead SNPs from the meta-analysis using the following formula :

$$\sum_{i=1}^{n} \frac{\beta_i^2 \cdot 2 \cdot EAF_i \cdot (1 - EAF_i)}{\sigma^2(residuals(\log(IL6)))}$$

where EAF is the effect allele frequency and the effect size of the individual variants and n is the total number of lead variants. The variance of the residuals of log(IL-6) was calculated using data from the NESDA cohort (N=2,517). The total common SNP heritability of serum IL-6 levels explained by all GWAS variants was estimated using the observed Z-statistics from the discovery analyses for a subset of pruned SNPs within our discovery association summary statistics. Following the original method (SumVg)[40], we pruned the imputed genotype dataset of the NESDA cohort using PLINK[41], removing highly correlated SNPs ($r^2>0.25$) within a 100-SNP sliding window and a step size of 25 SNPs per move. This resulted in a pruned set of 163,459 SNPs.

Fine mapping and identification of putative causal risk variants and candidate genes

Using 1000 Genomes sequence data (Phase1 Integrated Release, Version 3, 2012.04.30), we searched for variants in high LD ($r^2>0.8$) within a 1 Mb region on either side of the lead SNPs using tools available in Liftover[42], VCFtools[43], and PLINK[41]. We subsequently annotated these variants using ANNOVAR[44] with the RefSeq[45] database for variant function and genic residence or distance, and looked them up in the GWAS catalog[46] to identify genome-wide significant associations with other phenotypes.

RESULTS

A total of 52,654 individuals of European descent from 26 cohorts were included in the discovery GWAS meta-analysis with up to 2,835,074 autosomal SNPs passing quality control. Four cohorts, being ALSPAC, MONICA/KORA, NTR and SardiNIA, identified genome-wide significant associations in the *ABO* region, whereas none of the other 22 cohorts did, either individually or combined. These cohorts conditioned their results on their relevant top-SNP in *ABO*, the results of which were included in the discovery meta-analyses.

We identified 94 variants that were genome-wide significantly associated with IL-6 levels ($p < 5x10^{-8}$), representing two independent genetic loci on chromosomes 1q21 and 6p21, each having one statistically independent representative lead SNP. The minor alleles of rs4537545 on chromosome 1q21 ($p=8.39x10^{-85}$) and rs660895 on chromosome 6p21 ($p=1.80x10^{-9}$) showed increased ln(IL-6) concentrations of 0.091(±0.005) and 0.036(±0.006), respectively (Table 1). We next took the two independent variants (one from the 1q21 locus, one from the 6p21 locus) forward for replication testing, plus an additional set of seven statistically independent variants, representing 7 independent loci that showed suggestive association ($5×10^{-8} < p < 1×10^{-5}$) in the discovery analyses.

Replication analyses were performed in 12 additional independent cohorts encompassing up to 14,774 individuals, using a combination of in-silico and de-novo genotyping and following the same QC and statistical procedures as for the discovery phase. Both two independent genome-wide significant lead SNPs were replicated ($p<0.05$) and genome-wide significant in the combined analysis of discovery and replication samples ($p=1.20x10^{-122}$ for rs4537545 in 1q21 and $p=1.56x10^{-10}$ for rs660895 on 6p21, respectively). (Table 1).

After combining the discovery and replication analyses one additional variant, rs6734238 at chromosome 2q14, reached genome-wide significance ($p=1.84x10^{-11}$). The 2q14 and 6p21 loci are completely novel. In both discovery and replication association analyses effect sizes were generally consistent across individual studies for genome-wide significant variants, and we did not observe evidence of heterogeneity ($I^2<0.5$, Cochran's Q statistic >0.05). The three independently, genome-wide significantly associated SNPs combined explained approximately 1.06% of the variance in levels of IL-6 using data from the NESDA cohort, and by using the SumVg method[40], we estimated the percentage of phenotypic variance explained by all common variants to be 4.45%.

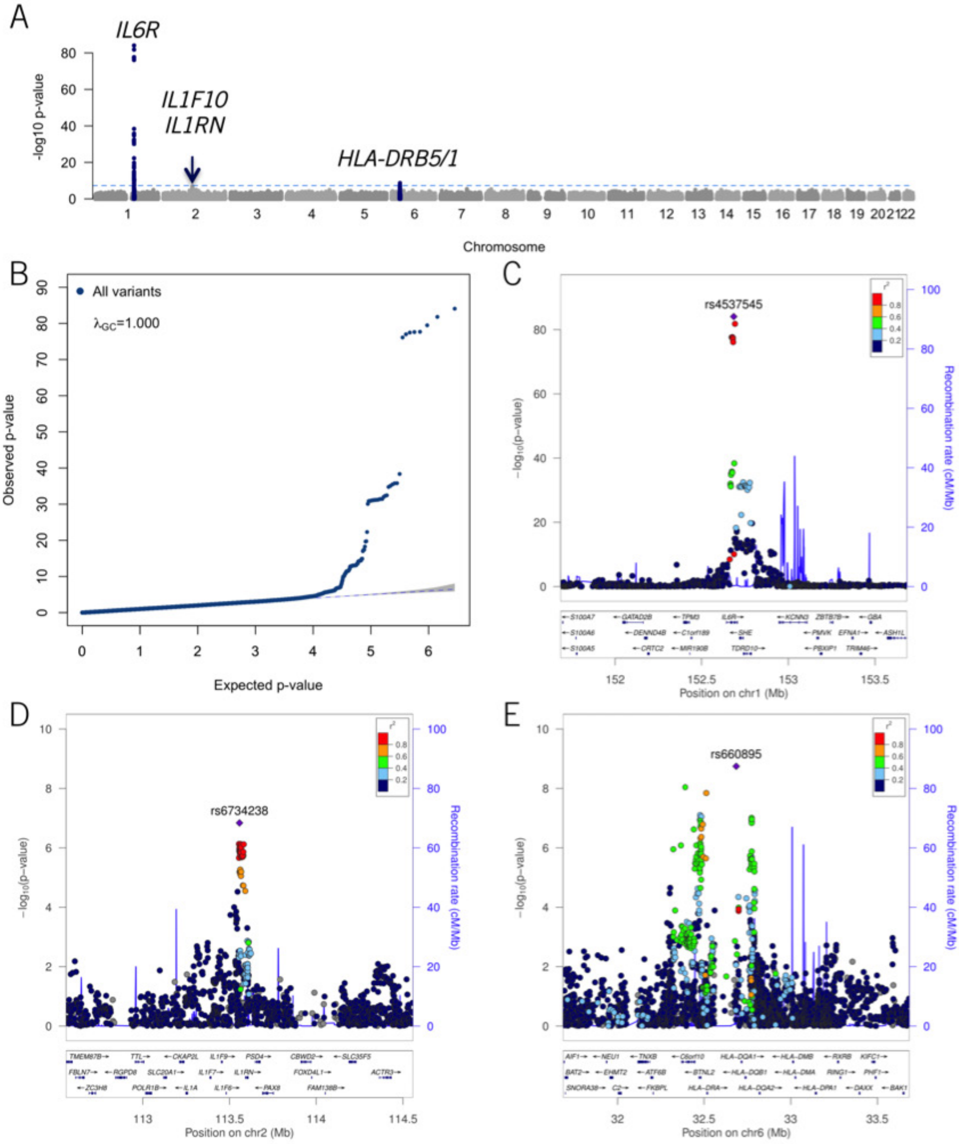


Figure 1. Manhattan plot and Locuszoom plots of the discovery analyses. a) Manhattan plot showing the association of SNPs with IL-6. Loci coloured in red or blue, three in total, represent those for which the lead SNPs reached genome-wide significance (P=5×10–8). Horizontal axis : relative genomic position of variants on the genome, vertical axis : -log10 p-value of each SNP; b) Quantile-quantile plot for p-values obtained from the meta-analysis. The horizontal and vertical axes represents the expected distribution of -log10(P-values) under the null hypothesis of no association, whereas the vertical axis shows the observed -log10(P-values). The blue dashed line represents the null, and λ_{gc} value represents the genomic inflation factor lambda. Each data point represents the observed versus the expected p-value of a variant included in the association analyses ; c-e) Regional association plots for each of the three genome-wide significant loci, 1q21, 2q14, and 6p21, respectively. Pairwise LD (r2) with the lead SNP is indicated following a color-coded scale. Horizontal axis : relative genomic position of variants within the locus, vertical axis : -log10 p-value of each SNP.

Table 1. Three independent variants in three genomic loci that were found to be genome-wide significantly associated with serum-levels of IL-6.

Locus	Variant	E/N	EAF	β(SE)	$P_{discovery}$	$P_{replication}$	$P_{combined}$	Function	Genes
1q21	rs4537545	T/C	0.39	0.091(0.005)	$8.39x10^{-85}$	$7.88x10^{-37}$	$1.20x10^{-122}$	intronic	*IL6R*
2q14	rs6734238	G/A	0.42	0.025(0.005)	$1.45x10^{-7}$	$3.24x10^{-5}$	$1.84x10^{-11}$	intergenic	*IL1F10,IL1RN*
6p21	rs660895	G/A	0.19	0.036(0.006)	$1.80x10^{-9}$	$3.38x10^{-2}$	$1.55x10^{-10}$	intergenic	*HLA-DRB5/1*

Variants are shown that reached $P < 5×10^{-8}$ in the combined analysis and are independent lead SNPs. Sample sizes: discovery cohorts, n=52,654; replication cohorts, n=14,774; combined, n=67,428. The effect sizes (β) in the discovery phase, given for the effect allele. E/N: E is the effect allele, and N is the non-effect allele. EAF: Effect Allele Frequency; Effect sizes and standard error (SE) values are in natural log (pg/ml) unit.

Fine mapping and identification of putative causal risk variants

The three identified loci harbour multiple immunologically associated genes ; lead variants were located within the Interleukin-6 Receptor (IL-6R, rs4537545, 1q21), in the vicinity of Interleukin-1 family member 10 and Interleukin 1 Receptor Antagonist (*IL1F10*,*IL1RN*, rs6734238, 2q14, intergenic), and near HLA class II histocompatibility antigen, DRB1 and DRB5 beta chain (*HLA-DRB1/DRB5*, rs660895, 6p21, intergenic). The search for functional variants in high LD (r^2 > 0.8) with the lead SNPs led to the identification of only one non-synonymous SNP, rs2228145, having an LD of r^2=0.95 with rs4537545, our lead SNP in the 1q21 locus.

A lookup of the three lead variants and their high-LD variants in the GWAS catalog (accessed October 23, 2015) revealed that for all of our loci one or more of the variants was associated with several diseases (asthma, rheumatoid arthritis, and IgA nephropathy) and serum levels of biomarkers (C-Reactive Protein (CRP), fibrinogen, and soluble Interleukin-6 receptor (sIL-6R)).

DISCUSSION

We have performed the largest meta-analysis of GWASs for IL-6 levels to date, that includes 66,341 samples from European descent. We found three independently associated loci, amongst which two novel, nearby or harbouring genes that have inflammatory roles.

The most strongly associated SNP, rs4537545, resides in the *IL-6R* gene, encoding the Interleukin-6 Receptor, and is in very high LD (r^2=0.95) with a non-synonymous SNP rs2228145 that results in an amino acid substitution at position 358 (Asp⟶Ala) on the extracellular domain of IL-6R. Many functional studies have been performed for this variant, showing that it impairs the responsiveness of cells targeted by IL-6[47],reduces *IL-6R* expression on cell surfaces[48], and increases levels of soluble IL-6R in individuals homozygous for this mutation[49,50] . Recently it has been demonstrated that increased levels of sIL-6R induced by this variant, can be explained by ectodomain shedding of IL-6R, a mechanism in which membrane-associated proteins are rapidly converted into soluble effectors whereby simultaneously cell surface expression of the same protein is reduced[51]. Increased levels of sIL-6R may act as a counter-balance to limit exaggerated IL-6 signalling, and may explain the protective effect of the 358Ala allele for various cardiovascular diseases including coronary artery disease (CAD)[52–54], atrial fibrillation (AF)[55], and abdominal aortic aneurysm (AAA)[56] as well as RA[57]. However in contrast with this finding, the IL-6-sIL-6R complex itself is capable of transducing IL-6 signalling to non-IL-6R expressing cells, known as trans-signalling[58], and it is this mechanism, as opposed to classic signalling, that is linked to chronic inflammatory disorders including IBD and RA[59]. Blocking IL-6 signalling cascades can be achieved by using an IL-6R specific inhibitor in the form of a monoclonal antibody, tocilizumab, which is a widely used therapy in the treatment of RA. Several variants in IL-6R, including rs2228145, may assist in the prediction of patient response to tocilizumab in rheumatoid arthritis[60]. The causal involvement of IL-6 levels in disease remains to be elucidated, but a recent study using a Mendelian randomisation (MR) approach did demonstrate that by using this SNP as instrumental variable, modelling the effects of tocilizumab, that IL-6R signalling has a causal effect on CAD[53]. On the other hand the pleiotropic nature of the IL-6R locus, influencing IL-6, CRP, and fibrinogen levels, prohibits instrumental variable analysis and attribution of causality to one particular intermediate.

Several other genes encompass the 1q21 locus, including Src Homology 2 Domain Containing E (*SHE*) and Tudor Domain containing 10 (*TDRD10*), and although we did not perform conditional analyses on the lead SNP in IL-6R, there are indications that no independent signals exist within this locus[25]. At the locus on chromosome 2, the lead SNP, rs6734238, is intergenic, but it has also been found to be associated with CRP[61,62] and fibrinogen[63]. The two nearest genes are Interleukin 1 Family Member 10 (*IL1F10*, distance=7,6 kB, currently known as IL-38) and Interleukin 1 Receptor Antagonist (*IL1RN*, distance=34,4 kB). *IL1F10* is one of several members of the interleukin 1 cytokine family that in concert are thought to regulate both innate and adapted immune responses. For IL-6 specifically it has been found that synthesis increases when dendritic cells are stimulated by bacterial lipopolysaccharides (LPS) in the presence of *IL1F10*[64]. IL-1RN is another member of the interleukin 1 cytokine family, with suggestive evidence for involvement in determining IL-6 levels in blood. One study found significant associations of one variant residing in this gene, rs4251961, with plasma CRP and IL-6 levels, albeit not independently replicated and not genome-wide significant ($P=1x10^{-4}$ and P=0.004, respectively)[65]. Our lead SNP was not in high LD ($r^2 > 0.8$) with variants in either neighbouring gene, and therefore in conjunction with its intergenic position identifying a causal variant in this locus remains non-trivial.

The 6p21 locus that was identified resides within the HLA region, which forms one of the most complex genomic regions to study due to its large LD blocks and sequence diversity. The nearest genes to our lead SNP constitute *HLA-DRB1* (distance=19,8 kB) and *HLA-DQA1* (distance=27,8 kB), both of which are histocompatibility complex genes that encode proteins that form complexes which are present on the surface of certain immune system cells that display fragments of foreign peptides to the immune system to trigger the body's immune response. This is the first time that variation near genes coding for antigen presenting complexes has been identified for inflammatory markers, even though our lead SNP in this locus has previously been found to be associated with diseases in which a dysfunctional immune system plays a major role. One high-LD variant (rs9272422, $r^2=0.82$ with our lead SNP, rs660895) residing in the promoter region of *HLA-DQA1*, confirms this role; it has been identified previously for Systemic Lupus Erythematosus (SLE)[66] and Ulcerative Colitis (UC)[67].

Various studies aimed to identify genetic variation underlying levels of IL-6[22–26] and found genome-wide significant associations in the *IL-6R* and *ABO* genes. The study performed by Shah and colleagues[25] found suggestive evidence (not genome-wide significant, best p-value in their respective study being $3.82x10^{-6}$) for additional loci, including *BUD13*, *TRIB3* and *SEZ6L*, none of which we could replicate here indicating that these might be false positives. A lookup of the variants identified by Shah et al. in the summary statistics of our discovery stage revealed no genome-wide significant variants in loci other than *IL-6R*, with the other variants having non-genome significant p-values, the highest being 0.65 (rs6139007 in *TRIB3*) and the lowest being 2.84x10^{-5} (rs4251961 in *IL1RN*, moderately in LD with our lead SNP rs6734238 in the 2q14 locus (r^2=0.576).

It is surprising that even with a sample size of 66,341, nearly eight times larger than the largest GWAS performed so far for IL-6 (N=8,356, Shah et al.), in total only three genetic loci (1q21, 2q14, and 6p21) could be identified, whereas we count the *ABO* locus as a likely artefact. Taken together the low estimates of explained levels of variance by our lead SNPs (~ 1%) and current estimates of explained heritability levels range between 15 to 61%, enormous increases in sample sizes would be required to identify additional variants explaining this missing heritability. Multiple explanations for this so-called missing heritability phenomenon have been proposed in the past[68], which can be sought in different classes of genetic variation such as rare variants[69], or can be explained by non-additive effects which may cause inflated estimates of heritability[70]. Plausible evidence for other sources of missing heritability that have been found are epigenetic changes[71], and haplotypes of common SNPs[72].

Collectively, our results provided additional insights into the biology of IL-6 synthesis. By substantially increasing sample size compared to previous studies we established two additional loci that are genome-wide significantly associated with IL-6 levels. In addition, we confirmed the known IL-6R locus and showed relevant biological mechanisms through which genetic variation or genes within our loci might contribute to determination of IL-6 levels.

Even though the strengths of our study are mainly sample size improvement and subsequent identification of additional loci, its main limitation is that mainly a by-now fairly limited set of mostly common variants have been investigated as genetic sources of variation in IL-6 levels. Future studies are recommended to aim to identify additional common genetic variation or rare variation with increasingly smaller effects, by firstly using deeper imputation panels, such as those of the UK10K project[73] or that of the Haplotype Reference Consortium, a strategy that holds great promise[69], and secondly by making use of genetically isolated populations[74]. Thirdly, we would like to stress the importance of phenotype harmonisation. As we identified genome-wide variants in the *ABO* locus, in four studies participating in the discovery, but not in the remaining 22 cohorts, there is a strong indication that this locus is assay-specific. Future collaborative efforts therefore should strive to use well-calibrated assays, standardised protocols for sample handling and processing[75], though this will be difficult to achieve in practice.

Web resources

QCGWAS, https://cran.r-project.org/web/packages/QCGWAS/index.html
GWAMA, http://www.well.ox.ac.uk/gwama/
METAL, http://csg.sph.umich.edu//abecasis/metal/
GCTA, http://www.complextraitgenomics.com/software/gcta/
LocusZoom, http://csg.sph.umich.edu/locuszoom/
1000 Genomes, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
VCFtools, http://vcftools.sourceforge.net/
ANNOVAR, http://www.openbioinformatics.org/annovar/

REFERENCES

1. Kishimoto, T. IL-6: from its discovery to clinical applications. Int. Immunol. 22, 347–352 (2010).
2. Nishimoto, N. & Kishimoto, T. Interleukin 6: from bench to bedside. Nat Clin Pract Rheum 2, 619–626 (2006).
3. Gelinas, L., Falkenham, A., Oxner, A., Sopel, M. & Légaré, J.-F. Highly purified human peripheral blood monocytes produce IL-6 but not TNFα in response to angiotensin II. Journal of Renin-Angiotensin-Aldosterone System 12, 295–303 (2011).
4. Kitani, A. et al. Autostimulatory effects of IL-6 on excessive B cell differentiation in patients with systemic lupus erythematosus: analysis of IL-6 production and IL-6R expression. Clin Exp Immunol 88, 75–83 (1992).
5. Li, T. & He, S. Induction of IL-6 release from human T cells by PAR-1 and PAR-2 agonists. Immunol Cell Biol 84, 461–466 (2006).
6. Ng, E. K. et al. Human intestinal epithelial and smooth muscle cells are potent producers of IL-6. Mediators Inflamm. 12, 3–8 (2003).
7. Fain, J. N. Release of interleukins and other inflammatory cytokines by human adipose tissue is enhanced in obesity and primarily due to the nonfat cells. Vitam. Horm. 74, 443–477 (2006).
8. Podor, T. J., Jirik, F. R., Loskutoff, D. J., Carson, D. A. & Lotz, M. Human endothelial cells produce IL-6. Lack of responses to exogenous IL-6. Ann. N. Y. Acad. Sci. 557, 374–385; discussion 386–387 (1989).
9. Sanchez, C., Gabay, O., Salvat, C., Henrotin, Y. E. & Berenbaum, F. Mechanical loading highly increases IL-6 production and decreases OPG expression by osteoblasts. Osteoarthritis and Cartilage 17, 473–481 (2009).
10. Scheller, J., Grötzinger, J. & Rose-John, S. Updating interleukin-6 classic- and trans-signaling. Signal Transduction 6, 240–259 (2006).
11. Madhok, R., Crilly, A., Watson, J. & Capell, H. A. Serum interleukin 6 levels in rheumatoid arthritis: correlations with clinical and laboratory indices of disease activity. Ann Rheum Dis 52, 232–234 (1993).
12. de Benedetti, F. et al. Correlation of Serum Interleukin-6 Levels with Joint Involvement and Thrombocytosis in Systemic Juvenile Rheumatoid Arthritis. Arthritis & Rheumatism 34, 1158–1163 (1991).
13. Cesari, M. et al. Inflammatory markers and onset of cardiovascular events: results from the Health ABC study. Circulation 108, 2317–2322 (2003).
14. Calabrese, L. H. & Rose-John, S. IL-6 biology: implications for clinical targeting in rheumatic disease. Nat Rev Rheumatol 10, 720–727 (2014).
15. Scheinecker, C., Smolen, J., Yasothan, U., Stoll, J. & Kirkpatrick, P. Tocilizumab. Nat Rev Drug Discov 8, 273–274 (2009).
16. Yazici, Y. et al. Efficacy of tocilizumab in patients with moderate to severe active rheumatoid arthritis and a previous inadequate response to disease-modifying antirheumatic drugs: the ROSE study. Ann. Rheum. Dis. 71, 198–205 (2012).
17. Yokota, S. et al. Efficacy and safety of tocilizumab in patients with systemic-onset juvenile idiopathic arthritis: a randomised, double-blind, placebo-controlled, withdrawal phase III trial. Lancet 371, 998–1006 (2008).
18. Tanaka, T., Ogata, A. & Narazaki, M. Tocilizumab for the treatment of rheumatoid arthritis. Expert Rev Clin Immunol 6, 843–854 (2010).

19. Wörns, M. A., Victor, A., Galle, P. R. & Höhler, T. Genetic and environmental contributions to plasma C-reactive protein and interleukin-6 levels--a study in twins. Genes Immun. 7, 600–605 (2006).

20. Sas, A. A. et al. The age-dependency of genetic and environmental influences on serum cytokine levels: a twin study. Cytokine 60, 108–113 (2012).

21. Neijts, M. et al. Genetic architecture of the pro-inflammatory state in an extended twin-family design. Twin Res Hum Genet 16, 931–940 (2013).

22. Melzer, D. et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 4, e1000072 (2008).

23. Naitza, S. et al. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. PLoS Genet. 8, e1002480 (2012).

24. Comuzzie, A. G. et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS ONE 7, e51954 (2012).

25. Shah, T. et al. Gene-centric analysis identifies variants associated with interleukin-6 levels and shared pathways with other inflammation markers. Circ Cardiovasc Genet 6, 163–170 (2013).

26. Sidore, C. et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat. Genet. 47, 1272–1281 (2015).

27. Dinse, G. E. et al. Accommodating measurements below a limit of detection: a novel application of Cox regression. Am. J. Epidemiol. 179, 1018–1024 (2014).

28. Uh, H.-W., Hartgers, F. C., Yazdanbakhsh, M. & Houwing-Duistermaat, J. J. Evaluation of regression methods when immunological measurements are constrained by detection limits. BMC Immunology 9, 59 (2008).

29. Vaez, A. et al. lodGWAS: a software package for genome-wide association anal-ysis of biomarkers with a limit of detection. Bioinformatics (2016). doi:10.1093/bioinformatics/btw021

30. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913 (2007).

31. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34, 816–834 (2010).

32. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).

33. Servin, B. & Stephens, M. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. PLoS Genet 3, e114 (2007).

34. van der Most, P. J. et al. QCGWAS: A flexible R package for automated quality control of genome-wide association results. Bioinformatics (2014). doi:10.1093/bioinformatics/btt745

35. Pei, Y.-F., Zhang, L., Li, J. & Deng, H.-W. Analyses and Comparison of Imputation-Based Association Methods. PLoS ONE 5, e10827 (2010).

36. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11, 288 (2010).

37. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42, 565–569 (2010).

38. Penninx, B. W. J. H. et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatr Res 17, 121–140 (2008).

39. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).

40. So, H.-C., Li, M. & Sham, P. C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. Genet. Epidemiol. 35, 447–456 (2011).

41. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).

42. Hinrichs, A. S. et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34, D590–598 (2006).

43. Danecek, P. et al. The variant call format and VCFtools. Bioinformatics 27, 2156–2158 (2011).

44. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).

45. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33, D501–504 (2005).

46. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42, D1001–1006 (2014).

47. Ferreira, R. C. et al. Functional *IL6R* 358Ala allele impairs classical IL-6 receptor signaling and influences risk of diverse inflammatory diseases. PLoS Genet. 9, e1003444 (2013).

48. Stone, K. et al. Interleukin-6 receptor polymorphism is prevalent in HIV-negative Castleman Disease and is associated with increased soluble interleukin-6 receptor levels. PLoS ONE 8, e54610 (2013).

49. Rafiq, S. et al. A common variant of the interleukin 6 receptor (IL-6r) gene increases IL-6r and IL-6 levels, without other inflammatory effects. Genes Immun 8, 552–559 (2007).

50. Galicia, J. C. et al. Polymorphisms in the IL-6 receptor (IL-6R) gene: strong evidence that serum levels of soluble IL-6R are genetically influenced. Genes Immun. 5, 513–516 (2004).

51. Hayashida, K., Bartlett, A. H., Chen, Y. & Park, P. W. Molecular and Cellular Mechanisms of Ectodomain Shedding. Anat Rec 293, 925–937 (2010).

52. *IL6R* Genetics Consortium Emerging Risk Factors Collaboration et al. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. Lancet 379, 1205–1213 (2012).

53. Interleukin-6 Receptor Mendelian Randomisation Analysis (*IL6R* MR) Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. Lancet 379, 1214–1224 (2012).

54. CARDIoGRAMplusC4D Consortium et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet. 45, 25–33 (2013).

55. Schnabel, R. B. et al. Large-Scale Candidate Gene Analysis in Whites and African Americans Identifies *IL6R* Polymorphism in Relation to Atrial Fibrillation The National Heart, Lung, and Blood Institute's Candidate

Gene Association Resource (CARe) Project. Circ Cardiovasc Genet 4, 557–564 (2011).

56. Harrison, S. C. et al. Interleukin-6 receptor pathways in abdominal aortic aneurysm. Eur. Heart J. 34, 3707–3716 (2013).

57. Eyre, S. et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet. 44, 1336–1340 (2012).

58. Scheller, J., Ohnesorge, N. & Rose-John, S. Interleukin-6 trans-signalling in chronic inflammation and cancer. Scand. J. Immunol. 63, 321–329 (2006).

59. Chalaris, A., Schmidt-Arras, D., Yamamoto, K. & Rose-John, S. Interleukin-6 trans-signaling and colonic cancer associated with inflammatory bowel disease. Dig Dis 30, 492–499 (2012).

60. Enevold, C. et al. Interleukin-6-receptor polymorphisms rs12083537, rs2228145, and rs4329505 as predictors of response to tocilizumab in rheumatoid arthritis. Pharmacogenet. Genomics 24, 401–405 (2014).

61. Dehghan, A. et al. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123, 731–738 (2011).

62. Reiner, A. P. et al. Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women. Am. J. Hum. Genet. 91, 502–512 (2012).

63. Sabater-Lleal, M. et al. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. Circulation 128, 1310–1324 (2013).

64. van de Veerdonk, F. L. et al. IL-38 binds to the IL-36 receptor and has biological effects on immune cells similar to IL-36 receptor antagonist. Proc. Natl. Acad. Sci. U.S.A. 109, 3001–3005 (2012).

65. Reiner, A. P. et al. Polymorphisms of the IL1-Receptor Antagonist Gene (*IL1RN*) Are Associated With Multiple Markers of Systemic Inflammation. Arterioscler Thromb Vasc Biol 28, 1407–1412 (2008).

66. Hom, G. et al. Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N. Engl. J. Med. 358, 900–909 (2008).

67. Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124 (2012).

68. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).

69. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. 47, 1114–1120 (2015).

70. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. PNAS 109, 1193–1198 (2012).

71. Cortijo, S. et al. Mapping the Epigenetic Basis of Complex Traits. Science 343, 1145–1148 (2014).

72. Bhatia, G. et al. Haplotypes of common SNPs can explain missing heritability of complex diseases. bioRxiv 022418 (2015). doi:10.1101/022418

73. UK10K Consortium et al. The UK10K project identifies rare variants in health and disease. Nature 526, 82–90 (2015).

74. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. Brief Funct Genomics 13, 371–377 (2014).

75. de Jager, W., Bourcier, K., Rijkers, G. T., Prakken, B. J. & Seyfert-Margolis, V. Prerequisites for cytokine measurements in clinical trials with multiplex immunoassays. BMC Immunol 10, 52 (2009).

SECTION I :
GWAS AND INFLAMMATORY MARKER GENETICS

5

Discovery and Fine Mapping of Serum Protein Loci through Transethnic Meta-analysis

American Journal of Human Genetics. 2012 Oct 5;91(4):744-53

Nora Franceschini,* Frank J.A. van Rooij*, Bram P. Prins*, Mary F. Feitosa*, Mahir Karakas*, John H. Eckfeldt, Aaron R. Folsom, Jeffrey Kopp, Ahmad Vaez, Jeanette S. Andrews, Jens Baumert, Vesna Boraska, Linda Broer, Caroline Hayward, Julius S. Ngwa, Yukinori Okada, Ozren Polasek, Harm-Jan Westra, Ying A. Wang, Fabiola Del Greco M., Nicole L. Glazer, Karen Kapur, Ido P. Kema, Lorna M. Lopez, Arne Schillert, Albert V. Smith, Cheryl A. Winkler, Lina Zgaga, The LifeLines Cohort Study, Stefania Bandinelli, Sven Bergmann, Mladen Boban, Murielle Bochud, Y.D. Chen, Gail Davies, Abbas Dehghan, Jingzhong Ding, Angela Doering, J. Peter Durda, Luigi Ferrucci, Oscar H. Franco, Lude Franke, Grog Gunjaca, Albert Hofman, Fang-Chi Hsu, Ivana Kolcic, Aldi Kraja, Michiaki Kubo, Karl J. Lackner, Lenore Launer, Laura R. Loehr, Guo Li, Christa Meisinger, Yusuke Nakamura, Christine Schwienbacher, John M. Starr, Atsushi Takahashi, Vesela Torlak, André G. Uitterlinden, Veronique Vitart, Melanie Waldenberger, Philipp S. Wild, Mirna Kirin, Tanja Zeller, Tatijana Zemunik, Qunyuan Zhang, Andreas Ziegler, Stefan Blankenberg, Eric Boerwinkle, Ingrid B. Borecki, Harry Campbell, Ian J. Deary, Timothy M. Frayling, Christian Gieger, Tamara B. Harris, Andrew A. Hicks, Wolfgang Koenig, Christopher J. O'Donnell, Caroline S. Fox, Peter P. Pramstaller, Bruce M. Psaty, Alex P. Reiner, Jerome I. Rotter, Igor Rudan, Harold Snieder, Toshihiro Tanaka, Cornelia M. van Duijn, Peter Vollenweider, Gerard Waeber, James F. Wilson, Jacqueline C.M. Witteman, Bruce H.R. Wolffenbuttel, Alan F. Wright, Qingyu Wu, Yongmei Liu, Nancy S. Jenny, Kari E. North,1 Janine F. Felix, Behrooz Z. Alizadeh, L. Adrienne Cupples, John R.B. Perry, and Andrew P. Morris

*Equal contribution

ABSTRACT

Many disorders are associated with altered serum protein concentrations, including malnutrition, cancer, and cardiovascular, kidney, and inflammatory diseases. Although these protein concentrations are highly heritable, relatively little is known about their underlying genetic determinants. Through transethnic meta-analysis of European-ancestry and Japanese genome-wide association studies, we identified six loci at genome-wide significance ($p < 5\times10^{-8}$) for serum albumin (*HPN-SCN1B*, *GCKR-FNDC4*, *SERPINF2-WDR81*, *TNFRSF11A-ZCCHC2*, *FRMD5-WDR76*, and *RPS11-FCGRT*, in up to 53,190 European-ancestry and 9,380 Japanese individuals) and three loci for total protein (*TNFRS13B*, 6q21.3, and *ELL2*, in up to 25,539 European-ancestry and 10,168 Japanese individuals). We observed little evidence of heterogeneity in allelic effects at these loci between groups of European and Japanese ancestry but obtained substantial improvements in the resolution of fine mapping of potential causal variants by leveraging transethnic differences in the distribution of linkage disequilibrium. We demonstrated a functional role for the most strongly associated serum albumin locus, *HPN*, for which *Hpn* knockout mice manifest low plasma albumin concentrations. Other loci associated with serum albumin harbor genes related to ribosome function, protein translation, and proteasomal degradation, whereas those associated with serum total protein include genes related to immune function. Our results highlight the advantages of transethnic meta-analysis for the discovery and fine mapping of complex trait loci and have provided initial insights into the underlying genetic architecture of serum protein concentrations and their association with human disease.

MAIN TEXT

Albumin, the major plasma protein, transports endogenous and exogenous compounds such as nutrients, hormones, metabolic catabolites, and drugs and maintains intravascular volume by generating oncotic pressure. Diverse conditions, including cancer, liver and kidney diseases, and acute and chronic inflammatory states, manifest reduced plasma albumin concentrations. Low plasma albumin is associated with increased risk of cardiovascular disease[1] and mortality[2]. Gamma globulins, the second most abundant type of plasma protein, are composed primarily of immunoglobulins (Ig), the effector arm of humoral immunity. Dysregulation of Ig may result from altered production in infectious and autoimmune diseases and in immunodeficiency syndromes and from increased loss in kidney disease[3]. To date, little is known about the genetic regulation of plasma proteins, and the pathophysiologic mechanisms leading to low albumin concentrations in many acute and chronic disease conditions remain obscure. Genetic tools may allow for the discovery of pathways in the metabolism, regulation, and/or disease processes associated with changes in these proteins and may provide insights into the immune system, cancer, inflammatory diseases, and malnutrition.

Serum albumin heritability estimates range from 0.36 to 0.77 in family and twin studies[4–7]. Recent genome-wide association studies (GWASs) of populations of eastern Asian ancestry have revealed genetic loci contributing to variation in blood protein concentrations: *GCKR* (MIM 600842*)-FNDC4* (MIM 611905) to serum albumin[8], *TNFRSF13B* (MIM 604907) to total protein[8], and *RPS11* (MIM 180471*)-FCGRT* (MIM 601437) to both traits[8,9]. These associations have not been previously examined in other ancestry groups, and much of the heritability of blood protein concentrations remains unexplained. To bridge this gap in our understanding of the genetic architecture of serum protein concentrations, we began by performing a meta-analysis of European-ancestry GWASs for albumin and total protein. Subsequently, we combined the European meta-analysis with data from a GWAS of Japanese ancestry with the aim of (1) identifying additional loci through increased sample size, (2) assessing the evidence of heterogeneity in allelic effects between ethnic groups, and (3) improving the resolution of fine mapping in associated regions by leveraging the expected differences in the structure of linkage disequilibrium (LD) between diverse populations.

The European-ancestry meta-analysis consisted of 53,190 individuals (from 20 GWASs) for serum albumin and 25,539 individuals (from six GWASs) for total protein (Tables S1–S3 and Supplemental Data available online). The procedures followed were approved by the institutional review board committees and are in accordance with the ethical standards of the institutional committees on human experimentation. All participants have given informed consent. Sample and SNP quality control (QC) were undertaken within each study. Each GWAS was then imputed at up to 2.5M autosomal SNPs with the use of CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) samples from phase II of the International HapMap Project[10]. Each SNP with minor allele frequency (MAF) >1% that passed QC was tested for association with serum albumin and total protein under an additive model after adjustment for study-specific covariates. The results of each GWAS were corrected for residual population structure using the genomic control inflation factor[11] and were combined via fixed-effect inverse-variance-weighted meta-analysis. The results of the meta-analysis were subsequently corrected by a second round of genomic control (λ_{GC} = 1.04 for serum albumin and λ_{GC} = 1.02 for total protein) to allow for population differences between studies.

The European-ancestry meta-analysis identified six genome-wide significant loci ($p < 5\times10^{-8}$) for serum albumin and two for total protein (Table 1, Figure S1). These included association signals for serum albumin at *HPN* (MIM 142440)-*SCN1B* (MIM 600235) ($p=3.3\times10^{-15}$), *SERPINF2* (MIM 613168)-*WDR81* (MIM 614218) ($p=6.8\times10^{-13}$), *TNFRSF11A* (MIM 603499)-*ZCCHC2* ($p=3.9\times10^{-9}$), and *FRMD5-WDR76* ($p=2.0\times10^{-8}$) and for total protein on chromosome 6q21.3 ($p=3.4\times10^{-9}$). We also confirmed the associations previously reported in eastern Asian populations at *GCKR-FNDC4* ($p=2.9\times10^{-14}$) and *RPS11-FCGRT* ($p=3.2\times10^{-8}$) for serum albumin and *TNFRSF13B* ($p=1.3\times10^{-10}$) for total protein. Inspection of the *HPN-SCN1B* locus (Figure 1) provided evidence for two independent association signals for serum albumin (rs4806073, $p=3.3\times10^{-15}$; rs11671010, $p=1.9\times10^{-13}$; CEU r^2 = 0.02, 4.3 kb apart). To perform conditional analyses at this locus, we applied genome-wide complex trait analysis (GCTA)[12] to the results of the European-ancestry meta-analysis and individual-level genotype data from the Atherosclerosis Risk in Communities Study (8,127 European American individuals, Table S1) and confirmed that both SNPs remained genome-wide significant after adjustment for the effect of the other (rs4806073, $p=1.6\times10^{-12}$; rs11671010, $p=1.5\times10^{-11}$).

The results of the European meta-analysis were then combined with a Japanese-ancestry GWAS (BioBank Japan Project) consisting of 9,380 individuals for serum albumin and 10,168 individuals for total protein (Tables S1–S3, Supplemental Data). Sample and SNP QC were undertaken within the BioBank Japan Project. The GWAS was imputed at up to 2.5 M autosomal SNPs with the use of Han Chinese in Beijing and Japanese in Tokyo (CHB+JPT) samples from phase II of the International HapMap Project[10]. Each SNP with MAF >1% that passed QC was then tested for association with serum albumin and total protein in the BioBank Japan Project under an additive model after adjustment for age and sex, and the results were corrected for residual population structure with the genomic control inflation factor (λ_{GC} = 0.98 for serum albumin and λ_{GC} = 1.08 for total protein). The European meta-analysis and the BioBank Japan Project GWAS were then combined via transethnic meta-analysis implemented with MANTRA (meta-analysis of transethnic association studies)[13].

Table 1. Loci Achieving Genome-wide Significance for Serum Albumin and Total Protein in European-Ancestry Populations.

Lead SNP	Chr	Position (Build 36)	EA	NEA	EAF	Beta	SE	pValue	Sample size	Locus
Serum Albumin										
rs4806073	19	40247030	C	T	0.93	0.0257	0.0033	$3.3x10^{-15}$	53,187	*HPN-SCN1B*
rs1260326	2	27584444	T	C	0.41	0.0124	0.0016	$2.9x10^{-14}$	53,189	*GCKR-FNDC4*
rs11078597	17	1565113	C	T	0.18	0.0205	0.0029	$6.8x10^{-13}$	38,231	*SERPINF2-WDR81*
rs13381710	18	58304309	G	A	0.3	0.0108	0.0018	$3.9x10^{-9}$	53,189	*TNFRSF11A-ZCCHC2*
rs16948098	15	42006899	A	G	0.06	0.0229	0.0041	$1.9x10^{-8}$	53,189	*FRMD5-WDR76*
rs739347	19	54693197	T	C	0.89	0.0186	0.0034	$3.2x10^{-8}$	38,231	*RPS11-FCGRT*
Total Protein										
rs3751991	17	16776011	A	C	0.11	0.0377	0.0059	$1.3x10^{-10}$	25,537	*TNFRSF13B*
rs204999	6	32217957	A	G	0.74	0.0251	0.0042	$3.4x10^{-9}$	25,537	*6q21.3*

Genome-wide significance is defined as $p < 5x10^{-8}$. The following abbreviations are used: Chr, chromosome; EA, effect-allele; NEA, non-effect allele; EAF, effect allele frequency.

This approach has the advantage of allowing for heterogeneity in allelic effects between ancestry groups by assigning studies to clusters according to a Bayesian partition model of similarity in terms of their allele frequency profile. Studies assigned to the same cluster

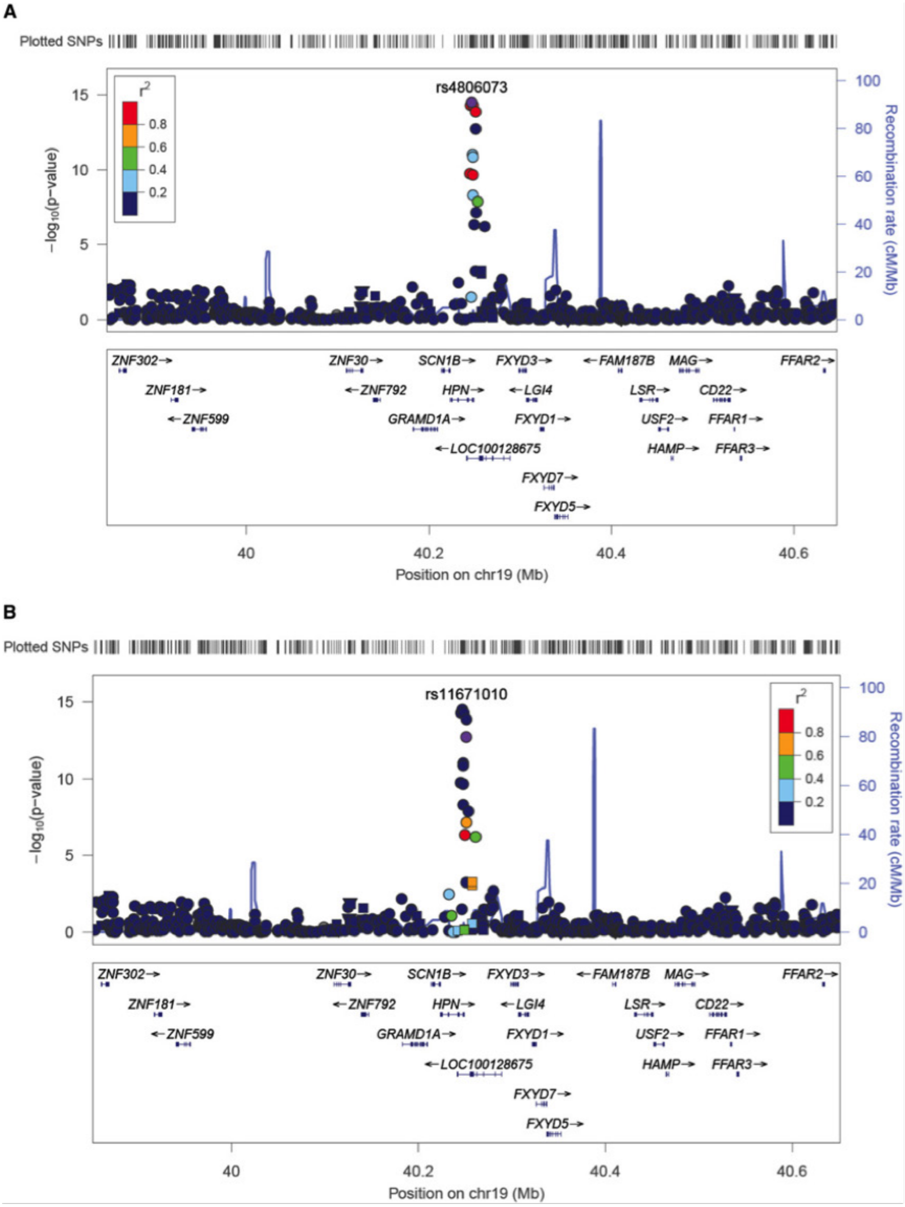


Figure 1. Signal Plot of the HPN-SCN1B Locus for Serum Albumin. The two panels present signal plots from the fixed-effect meta-analyses of European-ancestry individuals showing evidence for two independent associations in the region only: (A) SNPs tagged by rs4806073 and (B) SNPs tagged by rs11671010. In each panel, the lead SNP is represented by the purple circle. Each point represents a SNP plotted with their p (on a $\log_{10}$ scale) as a function of genomic position (build 36). The color coding of all other SNPs indicates LD with the lead SNP (estimated by CEU r^2 from phase II HapMap): red, $r^2 \geq 0.8$; gold, $0.6 \leq r^2 < 0.8$; green, $0.4 \leq r^2 < 0.6$; cyan, $0.^2 \leq r^2 < 0.4$; blue, $r2 < 0.^2$; and gray, r^2 unknown. Recombination rates are estimated from the International HapMap Project, and gene annotations are taken from the University of California Santa Cruz genome browser.

have the same allelic effect. However, each cluster can have different allelic effects. Fixed-effect meta-analysis is thus equivalent to a Bayesian partition model with a single cluster of studies.

We observed strong evidence of association, as defined by a $\log_{10}$ Bayes factor (BF) of 5 (equivalent to prior odds of association of any SNP with either trait of 1:100,000)[14], at all identified loci for both traits (Table 2, Figure S2). These loci included *RPS11-FCGRT* for total protein, which was previously observed in eastern Asian populations but not at genome-wide significance in our European-ancestry meta-analysis. The only exception was at the *FRMD5-WDR76* locus ($\log_{10}$BF = 4.79), where the lead SNP from the European meta-analysis (rs16948098) was not observed in the Japanese GWAS and is monomorphic in eastern Asian (CHB and JPT) HapMap populations[10]. Using the threshold of $\log_{10}$BF > 5 for strong evidence of association, we identified two additional "potential" loci for serum albumin and four for total protein (Table 2).

MANTRA revealed little evidence of heterogeneity in allelic effects between European-ancestry and Japanese studies at the majority of the serum albumin and total protein loci (Table 2). The extent of heterogeneity was assessed through comparison of association BF under a Bayesian partition model wherein the number of clusters of studies is unrestricted to that wherein there is a single cluster, the latter corresponding to homogeneous allelic effects across all ancestry groups. Subsequent fixed-effect inverse-variance-weighted meta-analysis across groups of European and Japanese ancestry (Table S4) revealed one additional signal for total-protein mapping to *ELL2* (MIM 601874, $p=1.1\times10^{-8}$), although none of the other potential MANTRA loci showed genome-wide significance ($p < 5\times10^{-8}$). Among these potential loci, however, there was strong evidence of heterogeneity at *ARID5B* (MIM 608538) for total protein (MANTRA $\log10_{BF}$ in favor of heterogeneity of 6.79). The lead SNP at this locus (rs2675609) was strongly associated with total protein only in the Japanese GWAS ($p=1.7\times10^{-6}$, compared with $p=0.014$ in the European meta-analysis), and the allelic effects were in opposite directions in the two ancestry groups (Table 2). Interestingly, the effect-allele frequency is similar in European-ancestry and Japanese GWASs, and there is little evidence of variation in LD structure between CEU and CHB+JPT reference haplotypes from the 1000 Genomes Project[15] (Figure S3). Although intrastudy phenotypic variation in total protein concentrations (such as Ig, which is not available for analyses here) might contribute to these apparent transethnic differences in allelic effects, further investigation is required to fully elucidate the source of heterogeneity between ancestry groups.

To assess the improvement in fine-mapping resolution due to transethnic meta-analysis in serum albumin and total protein loci, we defined "credible sets" of SNPs (J.B. Maller, personal communication) with the strongest signals of association and, hence, most likely to be causal (or tagging an unobserved causal variant), on the basis of European-ancestry GWASs only and then after inclusion of the Japanese study.
At each locus, defined by the genomic region 500 kb up and downstream of the lead SNP, we calculated the posterior probability that the jth SNP is "causal" (or tags an unobserved causal variant) by

$$\phi_j = \frac{BF_j}{\sum_k BF_k}.$$

In this expression, BF_j denotes the BF in favor of association of the *j*th SNP from the transethnic MANTRA analysis, and the summation in the denominator is over all SNPs passing QC across the locus (J.B. Maller, personal communication). A 100ω% credible set at the locus was then constructed through (1) ranking all SNPs according to their BF and (2) combining ranked SNPs until their cumulative posterior probability exceeded ω. Using this definition, we observed improved resolution, in terms of the number of SNPs and the genomic interval covered by the credible set, at *HPN-SCN1B*, *TNFRSF11A-ZCCHC2*, and *RPS11-FCGRT* for serum albumin and at *TNFRSF13B* and the 6q21.3 locus for total protein (Figure 2, Table S5). The most striking improvements in resolution were observed at the 6q21.3 locus for total protein, wherein the 99% credible set was reduced from 14 SNPs (covering 346 kb), to just three (covering 37 kb). Furthermore, after transethnic meta-analysis, the posterior probability that the lead SNP was causal (or tagged an unobserved causal variant) was more than 95% at *GCKR-FNDC4* and *SERPINF2-WDR81* for serum albumin and at *TNFRSF13B* and the 6q21.3 locus for total protein.

Two of the serum albumin loci, *HPN-SCN1B* and *RPS11-FCGRT*, can be validated by existing mouse models. The lead SNP at the *HPN-SCN1B* locus maps to an intron of *HPN*, a gene encoding hepsin, a membrane-bound serine protease that has substrate

Table 2.Loci with Strong Evidence of Association with Serum Albumin and Total Protein after Transethnic MANTRA Analysis of European-Ancestry and Japanese GWASs.

			Alleles		European Ancestry GWAS Meta-analysis					Japanese GWAS					MANTRA Transethnic Meta-analysis		
Lead SNP	Chr	Position (Build 36)	Effect	Other	EAF	Beta	SE	p Value	Sample Size	EAF	Beta	SE	p Value	Sample Size	$\log_{10}$ BF Association	$\log_{10}$ BF Heterogeneity	Locus
Serum Albumin Established Loci																	
rs1260326	2	27,584,444	T	C	0.41	0.0124	0.0016	$2.9x10^{-14}$	53,189	0.56	0.0270	0.0050	$2.2x10^{-8}$	9,380	17.01	0.28	*GCKR-FNDC4*
rs4806073	19	40,247,030	C	T	0.93	0.0257	0.0033	$3.3x10^{-15}$	53,187	0.92	0.0380	0.0090	$1.3x10^{-5}$	9,380	15.81	-0.08	*HPN-SCN1B*
rs11078597	17	1,565,113	C	T	0.18	0.0205	0.0029	$6.8x10^{-13}$	38,231	0.18	0.0200	0.0060	$1.8x10^{-3}$	9,380	12.51	0.20	*SERPINF2-WDR81*
rs694419	18	58,277,092	T	C	0.52	0.0093	0.0016	$1.2x10^{-8}$	53,189	0.05	0.0180	0.0110	$9.2x10^{-2}$	9,380	7.05	-0.11	*TNFRSF11A-ZCCHC2*
rs2280401	19	54,691,821	A	G	0.17	0.0121	0.0024	$7.6x10^{-7}$	53,189	0.16	0.0240	0.0070	$3.1x10^{-4}$	9,380	6.96	0.13	*RPS11-FCGRT*
rs16948098	15	42,006,899	A	C	0.06	0.0229	0.0041	$1.9x10^{-8}$	53,189	0.00					4.79		*FRMD5-WDR76*
Serum Albumin Additional Potential Loci																	
rs2293579	11	47,397,334	A	G	0.40	0.0093	0.0017	$8.0x10^{-8}$	53,189	0.26	0.0030	0.0050	$5.7x10^{-1}$	9,380	5.61	-0.05	*PSMC3*
rs12914385	15	76,685,778	C	T	0.61	0.0064	0.0016	$8.7x10^{-5}$	53,189	0.70	0.0200	0.0050	$1.3x10^{-4}$	9,380	5.07	0.12	*CHRNA3-CHRNA5*
Total Protein Established Loci																	
rs4561508	17	16,789,475	T	C	0.11	0.0360	0.0060	$1.3x10^{-9}$	25,534	0.37	0.0470	0.0700	$2.0x10^{-11}$	10,168	16.25	0	*TNFRSF13B*
rs204999	6	32,217,957	A	G	0.74	0.0250	0.0040	$3.4x10^{-9}$	25,537	0.94	0.0420	0.0130	$1.7x10^{-3}$	10,168	9.01	0.10	*6q21.3*
rs2280401	19	54,691,821	A	G	0.16	0.0120	0.0050	$1.5x10^{-2}$	25,537	0.16	0.0500	0.0090	$6.5x10^{-8}$	10,168	5.19	0.96	*RPS11-FCGRT*
Total Protein Additional Potential Loci																	
rs3777200	5	95,260,547	T	C	0.27	0.0180	0.0040	$1.9x10^{-5}$	25,524	0.30	0.0290	0.0070	$1.1x10^{-4}$	10,168	6.57	-0.14	*ELL2*
rs1260326	2	27,584,444	T	C	0.44	0.0150	0.0040	$1.1x10^{-4}$	25,537	0.56	0.0310	0.0070	$3.7x10^{-6}$	10,168	5.93	-0.02	*GCKR-FNDC4*
rs2675609	10	63,306,537	T	C	0.39	0.0090	0.0040	$1.4x10^{-2}$	25,537	0.43	-0.036	0.0070	$1.7x10^{-6}$	10,168	5.81	6.79	*ARID5B*
rs10097731	8	82,190,227	T	G	0.15	0.0230	0.0050	$1.1x10^{-5}$	25,537	0.17	0.0310	0.0090	$5.9x10^{-4}$	10,168	5.67	0.19	*PAG1*

Strong evidence is defined as log10 BF > 5. The following abbreviations are used: Chr, chromosome; GWAS, genome-wide association study; MANTRA, meta-analysis of transethnic association studies; EAF, effect allele frequency; and BF, Bayes factor.

specificity for basic amino acids similar to that of proalbumin processing, suggesting a physiologic role of hepsin in the cleavage of proalbumin to albumin. We compared serum protein concentrations between hepsin knockout (KO) mice and wild-type litter mates[16] (Figure 3) using blood samples collected from the inferior vena cava and analyzed by Consolidated Veterinary Diagnostics (West Sacramento, CA, USA). In *hepsin*$^{-/-}$ mice, we observed overwhelming evidence of reduced serum albumin ($p=9.1\times10^{-12}$) and, to a lesser extent, reduced total protein ($p=1.5\times10^{-5}$), but not Ig. At the *RPS11-FCGRT* locus, KO *Fcgrt* mice have been previously demonstrated to manifest low serum albumin and low serum gamma Ig concentrations[17,18].

Furthermore, in humans, two siblings with genetic deficiency of FcRn due to lack of the β2 microglobumin component have manifested reduced serum albumin and gamma Ig concentrations[19]. *FCGRT* encodes the heavy alpha chain of the FcRn, which prevents lysosomal degradation of albumin and Ig in lysosomes and thereby extends their serum half-life[17].

To gain insights into the possible functional role of other serum albumin and total protein loci, we began by performing expression quantitative trait locus (eQTL) mapping using data derived from 1,469 whole blood samples[20]. Gene expression levels were measured from peripheral blood and assayed in 1,240 individuals with Illumina HT12 v3 and in 229 individuals with Illumina H8 v2 BeadChip arrays. Both sets were independently quantile-normalized after $\log_2$ transformation and subsequently corrected for 50 principal components obtained from the gene expression probe covariance matrix.

To integrate both data sets, genotype data were imputed up to 2.5M autosomal SNPs with the use of CEU samples from phase II of the International HapMap Project.10 SNPs of low frequency (MAF < 5%) or with deviation from Hardy-Weinberg equilibrium ($p < 10^{-4}$) were excluded from the subsequent analysis. *Cis*-eQTL effects (within 1 Mb of the probe) were determined with Spearman's ranked correlation, and meta-analysis between the two data sets was performed with the use of weighted *Z* scores. The false discovery rate (FDR) was then assessed by permutation. Using this approach, we mapped lead SNPs at four of the identified loci (*GCKR-FNDC4*, *SERPINF2-WDR81*, and *RPS11-FCGRT* for serum albumin; 6q21.3 for total protein) to cis expression levels of 18 genes (Table S6). The strongest associations were observed for expression of *NOSIP* ($p=2.4\times10^{-17}$ at *RPS11-*

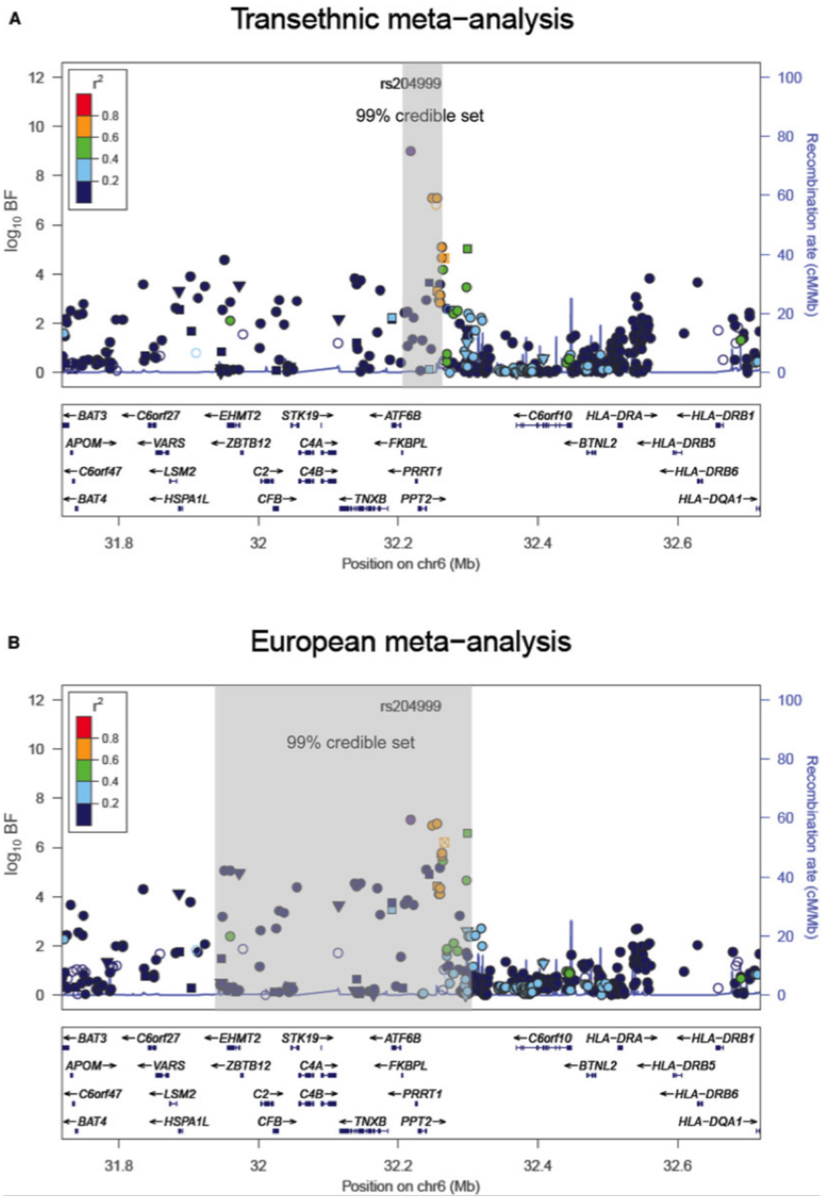


Figure 2. Fine Mapping of the 6q21.3 Locus for Total Protein. The two panels present signal plots for the MANTRA association signal (A) after transethnic meta-analysis of European-ancestry and Japanese GWASs and (B) after meta-analysis of the European ancestry GWAS only. Each point represents a SNP passing QC in our MANTRA analysis, plotted with their BF (on a log10 scale) as a function of genomic position (build 36). In each panel, the lead SNP is represented by the purple circle. The color coding of all other SNPs indicates LD with the lead SNP (estimated by CEU r2 from phase II HapMap): red, r2 ≥ 0.8; gold, 0.6 ≤ r2<0.8; green, 0.4 ≤ r2<0.6; cyan, 0.2 ≤ r2<0.4; blue, r2<0.2; and gray, r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser. In each panel, the gray-shaded regions correspond to the genomic interval covered by a 99% credible set of SNPs.

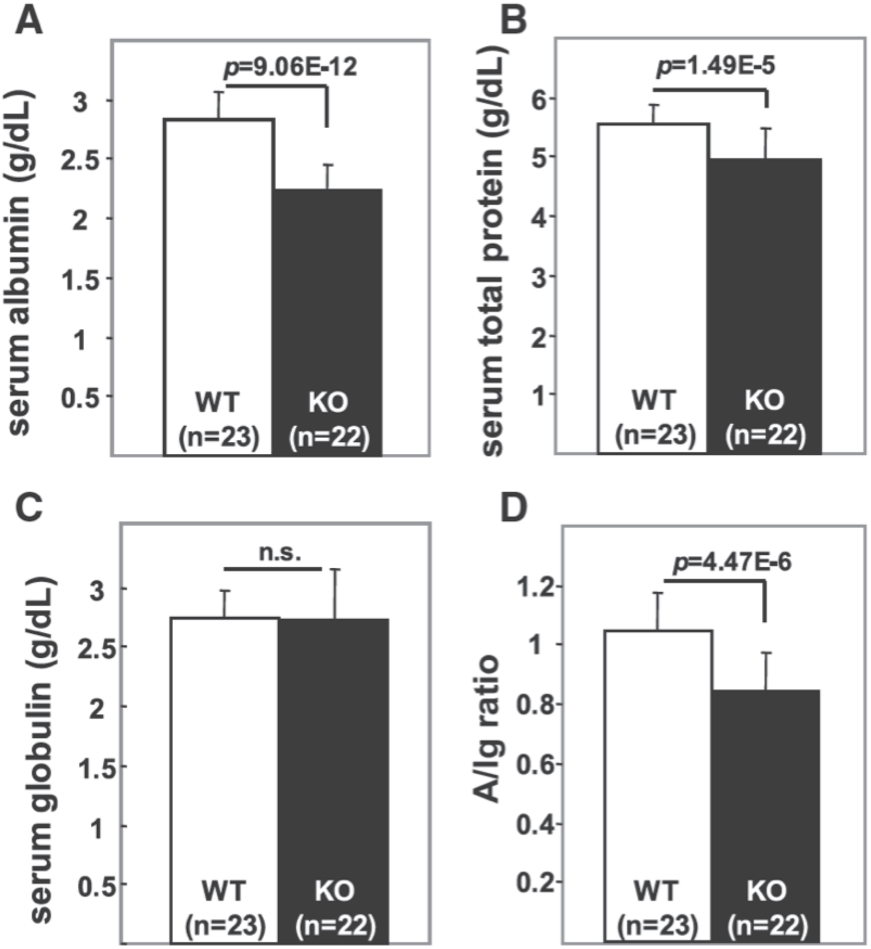


Figure 3. Serum Albumin, Total Protein, and Serum Globulin Concentrations and Albumin-to-Ig Ratio in Wild-Type and hepsin−/− Mice. Data are presented as mean ± SD, and the number of mice in each experimental group is shown in parentheses. Results for serum albumin are shown in (A); for total protein in (B); for serum globulin in (C); and the albumin-to-Ig ratio (A/Ig ratio) in wild-type (WT) and hepsin−/− (KO) mice in (D). Statistical differences are shown by p values. Note the significantly lower serum albumin concentrations, but not Ig concentrations, in KO mice compared to WT mice.

FCGRT) and *HLA-DQA1* (MIM 146880) / *HLA-DQA2* (MIM 613503, p=9.1×10^{-36} at 6q21.3). *NOSIP* (nitric oxide synthase interacting protein) inhibits endothelial nitric oxide synthesis, whereas *HLA-DQA1/2* is a human leukocyte antigen (HLA) class II antigen with an immune system role related to processing and presentation of antigen peptides.

We noticed that the lead SNP at the *GCKR-FNDC4* locus (rs1260326, c.1337T>C [p. Leu446Pro]; RefSeq NM_001486.3) is a *GCKR* missense mutation with moderate predicted functional impact by snpEff, and has been previously associated with several metabolic traits, as well as kidney, liver, and hematologic phenotypes (Table S7). We used data from the 1000 Genomes Project15 to search for additional coding variants with predicted function in strong LD ($r^2 > 0.5$ in 283 individuals of European ancestry) with lead SNPs at our identified serum albumin and total protein loci (Table S8). Across serum albumin loci, two nonsynonymous SNPs mapped to *SERPINF2*, a gene encoding an inhibitor of plasmin which degrades plasma fibrin and other proteins, and one nonsynonymous SNP mapped to *CHRNA5* (MIM 118505), a nicotinic acetylcholine receptor gene associated with smoking behaviour[21,22] and lung cancer[23]. For total protein, nonsynonymous SNPs mapped to *TNFRSF13B* and *ELL2*, and to *PPT2* (MIM 603298) and *EGFL8* (MIM 609897) at the 6q21.3 locus. Mutations in *TNFRSF13B* cause immunodeficiency common variable type 2 (MIM 240500), characterized by hypogammaglobulinemia and recurrent bacterial infections due to failure of β cell differentiation and impaired production of Ig. They also cause selective IgA deficiency 2 (MIM 609529), the most common primary immunodeficiency, affecting 1 in 600 individuals in the western world. *ELL2* product directs Ig secretion in plasma cells, and the 6q21.3 major histocompatibility complex class III region encompasses a number of genes involved in autoimmunity, inflammation, and complement proteins. Interrogation of the National Human Genome Research Institute (NHGRI) GWAS catalog[24] highlighted that lead SNPs at ten of the identified loci have themselves been reported or are in LD ($r^2 > 0.5$ in 283 individuals of European ancestry) with those disclosed, for a diverse range of human complex traits (Table S7) but are enriched for metabolic phenotypes that are associated with, or are direct products of, protein metabolism.

Finally, we used the human interactome database (Cytoscape) to construct an interaction network consisting of 250 proteins that directly interact with genes in the identified serum albumin and total protein loci reported in Table 2. For identifying molecular complexes within this first-degree interaction network, cluster analyses were performed with the FAG-EC algorithm, implemented in the ClusterViz plug-in, with standard settings applied. In total, 16 distinct clusters were identified, including three large complexes (Figure S4) that were carried forward for further analysis. Functional enrichment analyses within these clusters were performed following defined pathways from BioCarta, KEGG, PANTHER, and Reactome via the Database for Annotation, Visualization, and Integrated Discovery (DAVID). The most significantly enriched clusters from protein interaction network analyses incorporated ribosomal functioning and protein translation, proteasomal protein degradation, and immune-response signaling (Table S9). As a complementary approach, we applied an implementation of gene-set enrichment analysis (MAGENTA[25]) to identify whether defined biological pathways from BioCarta, Gene Ontology, Ingenuity, KEGG, PANTHER, and Reactome were enriched in the leading-edge fraction of the meta-analysis. In brief, gene association p values were calculated on the basis of meta-analysis summary statistics for SNPs within a 110-kb-upstream and 40-kb-downstream window. These gene scores were then corrected for gene size, number of SNPs, and the LD between them, and subsequently ranked by p value. Enrichment in the 75th and 95th percentiles was assessed for significance by comparison with 10,000 randomly generated pathways. Using an FDR threshold of 5%, we observed significant overrepresentation of genes assigned to three pathways: RNA- (FDR = 0.044), sensory-perception- (FDR = 0.027), and protein-trafficking-related pathways (FDR = 0.043-0.044) (Table S10).

In conclusion, we have identified six loci for serum albumin concentration and three for total protein at genome-wide significance. These loci harbor genes that fall across a diverse range of biological pathways, including those involved in biomarkers, immune regulation, and disease, but are enriched for those relevant to the synthesis and degradation of serum protein. By combining GWAS data from European and Japanese populations, we observed some evidence of heterogeneity in allelic effects between ancestry groups and have demonstrated substantial improvements in the localization of potential causal variants. Taken together, our results highlight the advantages of transethnic meta-analysis for the discovery and fine mapping of complex trait loci and provide initial insights into the underlying genetic architecture of serum protein concentrations and their association with human disease.

Acknowledgements

A.P.M acknowledges financial support from the Wellcome Trust, grant numbers WT081682, WT064890, WT098017, and WT090532. N.F. acknowledges financial support from AHA 0675001N, R01HL089651, and U01HG004803.

Supplementary material

Supplementary Material is available at the American Journal of Human Genetics online.

Web resources

The URLs for data presented herein are as follows:

1000 Genomes Project, http://www.1000genomes.org/
ClusterViz plug-in, http://code.google.com/p/clusterviz-cytoscape
DAVID, http://david.abcc.ncifcrf.gov
GCTA, http://www.complextraitgenomics.com/software/gcta/
NHGRI Catalog of Published Genome-wide Association Studies, http://www.genome.gov/gwastudies/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
SNIPPER, http://csg.sph.umich.edu/boehnke/snipper/
snpEff, http://snpeff.sourceforge.net

REFERENCES

1. Nelson, J. J. et al. Serum albumin level as a predictor of incident coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) study. Am. J. Epidemiol. 151, 468–477 (2000).
2. Goldwasser, P. & Feldman, J. Association of serum albumin and mortality risk. J Clin Epidemiol 50, 693–703 (1997).
3. Tietz Textbook of Clinical Chemistry. (Saunders, 1999).
4. Dal Colletto, G. M., Krieger, H. & Magalhães, J. R. Genetic and environmental determinants of 17 serum biochemical traits in Brazilian twins. Acta Genet Med Gemellol (Roma) 32, 23–29 (1983).
5. Whitfield, J. B. & Martin, N. G. The effects of inheritance on constituents of plasma: a twin study on some biochemical variables. Ann. Clin. Biochem. 21 (Pt 3), 176–183 (1984).
6. Kalousdian, S., Fabsitz, R., Havlik, R., Christian, J. & Rosenman, R. Heritability of clinical chemistries in an older twin cohort: the NHLBI Twin Study. Genet. Epidemiol. 4, 1–11 (1987).
7. Pankow, J. S. et al. Familial and genetic determinants of systemic markers of inflammation: the NHLBI family heart study. Atherosclerosis 154, 681–689 (2001).
8. Kamatani, Y. et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat. Genet. 42, 210–215 (2010).
9. Kim, Y. J. et al. Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. Nat. Genet. 43, 990–995 (2011).
10. International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861 (2007).
11. Devlin, B. & Roeder, K. Genomic control for association studies. Biometrics 55, 997–1004 (1999).
12. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82 (2011).
13. Morris, A. P. Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. 35, 809–822 (2011).
14. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. 10, 681–690 (2009).
15. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).
16. Wu, Q. et al. Generation and characterization of mice deficient in hepsin, a hepatic transmembrane serine protease. J. Clin. Invest. 101, 321–326 (1998).
17. Chaudhury, C. et al. The major histocompatibility complex-related Fc receptor for IgG (FcRn) binds albumin and prolongs its lifespan. J. Exp. Med. 197, 315–322 (2003).
18. Roopenian, D. C. et al. The MHC class I-like IgG receptor controls perinatal IgG transport, IgG homeostasis, and fate of IgG-Fc-coupled drugs. J. Immunol. 170, 3528–3533 (2003).
19. Wani, M. A. et al. Familial hypercatabolic hypoproteinemia caused by deficiency of the neonatal Fc receptor, FcRn, due to a mutant beta2-microglobulin gene. Proc. Natl. Acad. Sci. U.S.A. 103, 5084–5089 (2006).
20. Fehrmann, R. S. N. et al. Trans-eQTLs

reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 7, e1002197 (2011).

21. Thorgeirsson, T. E. et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat. Genet. 42, 448–453 (2010).

22. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat. Genet. 42, 441–447 (2010).

23. Amos, C. I. et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet. 40, 616–622 (2008).

24. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. 106, 9362–9367 (2009).

25. Segrè, A. V. et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 6, (2010).

SECTION II:

INTEGRATIVE POST-GWAS ANALYSES AND SYSTEMS GENETICS

6

In Silico Post Genome-Wide Association Studies Analysis of C-Reactive Protein Loci Suggests an Important Role for Interferons

Circulation Cardiovascular Genetics. 2015 Jun;8(3):487-97

Vaez A, Jansen R*, Prins BP*, Hottenga JJ, de Geus EJ, Boomsma DI, Penninx BW, Nolte IM, Snieder H, Alizadeh BZ

*Equal contribution

ABSTRACT

Background

Genome-wide association studies (GWASs) have successfully identified several single nucleotide polymorphisms (SNPs) associated with serum levels of C-reactive protein (CRP). An important limitation of GWASs is that the identified variants merely flag the nearby genomic region and do not necessarily provide a direct link to the biological mechanisms underlying their corresponding phenotype. Here we apply a bioinformatics-based approach to uncover the functional characteristics of the 18 SNPs that had previously been associated with CRP at a genome-wide significant level.

Methods and Results

In the first phase of in silico sequencing, we explore the vicinity of GWAS SNPs to identify all linked variants. In the second phase of expression quantitative trait loci analysis, we attempt to identify all nearby genes whose expression levels are associated with the corresponding GWAS SNPs. These 2 phases generate several relevant genes that serve as input to the next phase of functional network analysis. Our in silico sequencing analysis using 1000 Genomes Project data identified 7 nonsynonymous SNPs, which are in moderate to high linkage disequilibrium (r^2>0.5) with the GWAS SNPs. Our expression quantitative trait loci analysis, which was based on one of the largest single data sets of genome-wide expression probes (n>5000) identified 23 significantly associated expression probes belonging to 15 genes (false discovery rate <0.01). The final phase of functional network analysis revealed 93 significantly enriched biological processes (false discovery rate <0.01).

Conclusions

Our post-GWAS analysis of CRP GWAS SNPs confirmed the previously known overlap between CRP and lipids biology. Additionally, it suggested an important role for interferons in the metabolism of CRP.

INTRODUCTION

C-reactive protein (CRP), a pentameric molecule, is the most widely studied inflammatory marker[1]. Elevated levels of serum CRP have been associated with increased risks of cancer[2], type 2 diabetes mellitus[3], hypertension[4], coronary heart disease[5], stroke[6], bipolar disorder[7], and overall mortality[8]. However, its causal contribution to the pathophysiology of chronic diseases remains controversial[9–12]. Serum levels of CRP are regulated by both genetic and environmental factors[11,13]. 1Ts heritability has been reported to range from 10% to 65%[14–17], and genome-wide association studies (GWASs) have successfully identified several genetic variants associated with CRP levels[12,18]. A recent meta-analysis of 25 GWAS studies, including >80 000 subjects, identified 18 CRP genetic variants at genome-wide significance[19].

One important limitation of GWASs is that the identified single nucleotide polymorphisms (SNPs) are not necessarily causally related to their associated traits or diseases. Many GWAS SNPs merely flag causal variants in their vicinity[20]. Hence, identifying associated SNPs by GWAS does not necessarily provide sufficient information on the biological mechanisms or pathways underlying their corresponding phenotype. Therefore, after a successful GWAS study, it is essential to perform additional post-GWAS analyses to translate the GWAS findings represented by index SNPs into biological knowledge[21]. For example, we previously demonstrated that serum protein levels are regulated by ribosomal functioning, proteasomal degradation, and immune-response signaling pathways, leading to a better functional understanding of the GWAS findings for serum protein levels[22]. However, an in-depth post-GWAS analysis for CRP variants has not yet been performed[19], which means that CRP GWAS findings have been insufficiently translated into biological function. Consequently, the gain in knowledge on underlying mechanisms controlling CRP level has been limited. Given the clinical relevance of CRP as an established biomarker for many complex chronic disorders, an extended post-GWAS analysis of CRP variants may unravel new mechanisms, which will improve our understanding of the metabolism of CRP and its relevance to disease pathology.

Here we applied a bioinformatics-based approach to uncover the functional characteristics of the 18 CRP-associated variants[19]. We first performed an in silico sequencing analysis using 1000 Genomes Project data[23] to identify nearby nonsynonymous coding variants.

Second, we performed an expression quantitative trait loci (eQTL) analysis using a large data set of blood expression probes to find regulatory variants. Third, we integrated the findings of the abovementioned phases by performing a functional network analysis to unravel the underlying biological processes.

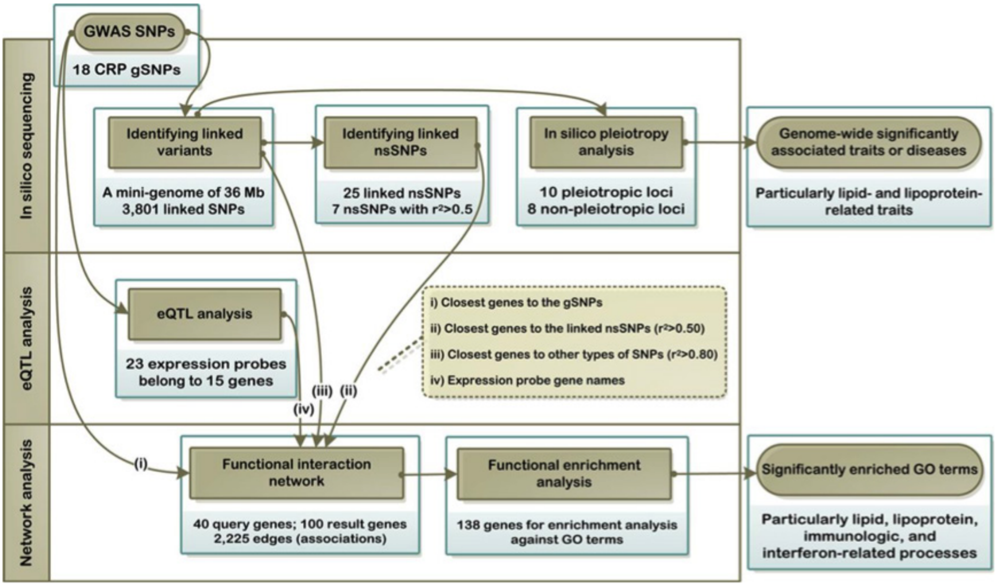

In silico sequencing
eQTL analysis
Network analysis
GWAS SNPs
18 CRP gSNPs
Identifying linked variants
A mini-genome of 36 Mb
3,801 linked SNPs
Identifying linked nsSNPs
25 linked nsSNPs
7 nsSNPs with r^2>0.5
In silico pleiotropy analysis
10 pleiotropic loci
8 non-pleiotropic loci
Genome-wide significantly associated traits or diseases
Particularly lipid- and lipoprotein-related traits
eQTL analysis
23 expression probes belong to 15 genes
i) Closest genes to the gSNPs
ii) Closest genes to the linked nsSNPs (r^2>0.50)
iii) Closest genes to other types of SNPs (r^2>0.80)
iv) Expression probe gene names
(iii) (ii)
(iv)
(i)
Functional interaction network
40 query genes; 100 result genes
2,225 edges (associations)
Functional enrichment analysis
138 genes for enrichment analysis against GO terms
Significantly enriched GO terms
Particularly lipid, lipoprotein, immunologic, and interferon-related processes

Figure 1. Flow diagram of the steps of CRP post-GWAS analysis. The inner grey boxes show the methods of the analysis, whereas the outer blue boxes show the main results of post-GWAS analysis of 18 genome-wide significantly associated CRP SNPs.

METHODS

We followed a bioinformatics-based approach, including 3 distinct phases, each consisting of multiple steps as described later (Figure 1).

Phase I: In Silico Sequencing

Identifying Linked Variants

First, we converted the chromosome positions of the GWAS SNPs (gSNPs) from the National Center for Biotechnology Information Build 36 (Human Genome 18) to National Center for Biotechnology Information Build 37 (Human Genome 19) using the LiftOver tool from the University of California Santa Cruz (UCSC) Genome Project[24]. Then, we targeted regions of 1 Mb at either side of each gSNP, resulting in a mini-genome of 36 Mb. The appropriate Variant Call Format[25] file for each 2 Mb region was downloaded from the 1000 Genomes Project ftp server using the Tabix software package[26]. We used the data from the 1000 Genomes Project Full Phase 1, November 2010 release (using August 2010 alignments), including only the 283 subjects of European ancestry[23]. Subsequently, for each Variant Call Format file, the r^2 between the gSNP and all other biallelic SNPs residing within the corresponding 2 Mb area was calculated as a metric of linkage disequilibrium (LD) using VCFtools[25]. Only those SNPs in moderate to high (r^2>0.50) LD with the corresponding gSNP were used in the next step of the analysis (Figure 1).

Identifying Linked Nonsynonymous SNPs

All these SNPs in LD with any of the gSNPs were annotated by ANNOVAR software[27] and then filtered in a stepwise manner. First, the SNPs were annotated to distinguish exonic variants from other variant types (intronic, intergenic, etc.). Nonexonic variants were excluded from further analyses. The remaining SNPs were annotated again to distinguish synonymous from nonsynonymous exonic SNPs, and synonymous SNPs were excluded. As a further step, the nonsynonymous SNPs (nsSNPs) were then characterized for their damaging effect on the corresponding protein using Sorting Intolerant From Tolerant (SIFT)[28] and Polymorphism Phenotyping (PolyPhen)[29] prediction scores. Their scores were obtained from Ensembl release 71 (accessed June 8, 2013)[30]. Whenever multiple scores were available for a single nsSNP, we selected the most damaging prediction scores as the smallest SIFT and the largest PolyPhen scores. These scores are just provided as Data Supplement about linked variants and hence, were not used in the downstream analyses.

In Silico Pleiotropy Analysis

To extend our knowledge of the possible function of the 18 CRP-associated loci, we sought to identify any trait or outcome associated with these 18 loci. Thus, for all gSNPs, as well as all SNPs in LD (r^2>0.80) with any of the gSNPs, we checked for genome-wide significant ($P<5\times10^{-8}$) pleiotropic effects on other complex traits or diseases identified in previous GWAS studies as listed in the National Human Genome Research Institute GWAS Catalog (Catalog of Published Genome-Wide Association Studies)[31] using ANNOVAR software (accessed June 13, 2013)[27]. However, as shown in Figure 1, the results of this step were not used in the downstream analyses, but were indeed used in the final interpretation of the results.

Phase II: eQTL Analysis

The data set of genome-wide expression probes and gene expression measurements have been described in more detail elsewhere[32,33].

Subjects

The 2 parent projects that supplied data for the eQTL analysis are large-scale longitudinal studies: the Netherlands Study of Depression and Anxiety[34] and the Netherlands Twin Registry[35]. The Netherlands Study of Depression and Anxiety and the Netherlands Twin Registry studies were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam, and all subjects provided written informed consent. The sample used for eQTL analysis after quality control consisted of 5071 subjects, 3109 from the Netherlands Twin Registry (from 1571 families: 614 dizygotic twin pairs, 1 monozygotic triplet, 668 monozygotic twin pairs, 394 siblings, and 148 unrelated subjects), and 1962 the Netherlands Study of Depression and Anxiety participants. The age of the participants ranged from 17 to 88 years (mean 38, SD 13), and 65% of the sample was female[32].

Blood Sampling, RNA Extraction, and Measurements

Venous blood samples were drawn in the morning after an overnight fast. Heparinized whole blood samples were transferred within 20 minutes of sampling into PAXgene Blood RNA tubes (Qiagen) and stored at −20 °C. Gene expression assays were conducted at the Rutgers University Cell and DNA Repository (http://www.rucdr.org). Samples were hybridized to Affymetrix U219 arrays containing 530 467 probes summarized in 49 293 probesets. All probes are 25 bases in length and designed to be perfect match complements to a designated transcript. Array hybridization, washing, staining, and scanning were performed in an Affymetrix GeneTitan System per the manufacturer’s protocol. Gene expression data were required to pass standard Affymetrix quality control metrics (Affymetrix expression console) before further analysis. Probes that did not map uniquely to Human Genome 19 or that contained a polymorphic SNP (dbSNP137 common with minor allele frequency >0.01) were removed for downstream analysis, resulting in 423 201 probes, summarized in 44 241 probesets, targeting 18 238 unique genes. Probeset expression values were obtained using robust multiarray average normalization implemented in Affymetrix Power Tools (APT, v1.12.0). Samples with low average correlation with other samples and samples with incorrect sex-chromosome expression were removed[32].

Genotype Data

DNA extraction has been described earlier[36]. Genotyping was done on multiple chip platforms for several partly overlapping subsets of participants. The following platforms were used: Affymetrix Perlegen 5.0, Illumina 370, Illumina 660, Illumina Omni Express 1 mol/L, and Affymetrix 6.0. After array-specific data analysis, genotype calls were made with the platform-specific software (Genotyper, Beadstudio). The extensive genotyping quality control steps and 1000 Genomes imputation procedures are described in the Data Supplement Text S1. Genotypes were coded into dosage format and filtered at minor allele frequency >0.01 and imputation quality of R^2>0.30 for eQTL analysis.

eQTL Analysis

Inverse quantile normal transformation was applied to the individual probeset data to obtain normal distributions. The transformed probeset data were then residualized with respect to the covariates sex, age, body mass index, smoking status, several technical covariates, and 3 principal components (PCs) from the genotype data. Genotype PCs were constructed using pruned GWAS data after removing ethnic outliers as described earlier[37]. The residualized probeset data were subjected to a principal component analysis to remove the first 50 PCs to adjust the gene expression levels for nongenetic variation, as proposed by Fehrmann et al. They have shown that removing expression PCs drastically increases the number of eQTLs[38]. We observe the same phenomenon in our data. Removing expression PCs has become a standard procedure in many eQTL studies[33,39]. Probesets at <1 Mb distance from the gSNPs were selected for eQTL analysis as follows: for each probeset–gSNP combination at maximally 1 Mb distance, a linear mixed model was fitted with expression level as dependent variable, genotype as fixed effect, and family ID and zygosity as random effects to account for family and twin relations[40]. Mixed models and resulting P values were computed using the function lmer from the lme4 R package (http://CRAN.R-project.org/package=lme4). To correct for multiple testing, false discovery rate (FDR) was computed using all P values from each probeset–gSNP combination at maximally 1 Mb distance using the function p.adjust from the stats R package, and any signal with FDR<0.01 was considered significant. The appropriate gene names of those significantly associated expression probes were then used in the next step as a set of prioritized biological candidate genes (Figure 1).

As a further step, for each locus with significant eQTL signal of FDR<0.01, we also identified the most significantly associated eQTL SNP (eSNP) for the corresponding transcript. We then performed conditional analyses to see if the gSNP is independently associated with the expression level. For conditional eQTL analysis, the transformed probeset data were residualized with respect to the corresponding eSNP before applying the mixed model. These eSNPs were not used in the downstream analysis (Figure 1).

Phase III: Network Analysis

Functional Interaction Network

To construct a functional association interaction network, we applied the GeneMANIA algorithm together with its large set of accompanying functional association data on coexpression, physical interaction, genetic interaction, shared protein domains, colocalization, and predicted association networks. This data set comprises 286 extended association networks[41].

We combined 4 biologically prioritized candidate gene sets into a single query gene set, which was used as input for the interaction network analysis: (1) closest genes to the gSNPs, (2) closest genes to the nsSNPs in high LD (r^2>0.50) with the corresponding gSNP, (3) closest genes to other types of SNPs in very high LD (r^2 >0.80) with the corresponding gSNP, and (4) expression probe gene names significantly (FDR<0.01) associated with gSNPs based on the eQTL analysis (Figure 1). We used different LD thresholds for nsSNPs than other types of SNPs as nsSNPs are more likely to be functionally important and also are more likely to reside within a lower frequency spectrum. Consequently, nsSNPs may be in modest LD with common gSNPs. Therefore, we used a more lenient LD threshold for nsSNPs (r^2>0.50) to ensure not to miss potentially functional variants with modest frequency and a standard LD threshold of r^2>0.80 for other types of SNPs.

Next, we constructed a weighted composite functional association network using the Cytoscape software platform[42], extended by the GeneMANIA plugin[43]. We selected all available networks option with a 100-gene output (accessed July 15, 2013).

Functional Enrichment Analysis

All the genes in the composite network, either from the query or the resulting gene sets, were then used for functional enrichment analysis against Gene Ontology terms (GO terms) to identify the most relevant GO terms using the same plugin[43]. Each GO annotation has an evidence code indicating the type of experimental or computational support for that association, for example, inferred from reviewed computational analysis (RCA) or inferred from electronic annotation (IEA). The first one (RCA) points to those predictions based on computational analyses of experimental data sets like protein–protein interaction or expression data. The latter (IEA) points to computationally assigned evidence codes, which have not been reviewed by a curator to verify their accuracy (http://www.geneontology.org/GO.evidence.shtml)[44]. IEA is the least reliable, but the most prevalent evidence code, that is, about 47% of all of the human GO annotations are based on IEA codes (accessed July 26, 2013). As both RCA and IEA annotations are solely based on computational predictions, the functional enrichment analysis was only performed against GO term annotations with non-IEA and non-RCA evidence codes to avoid circularity[44]. We considered any GO term with FDR <0.01 as significantly and those GO terms with FDR between 0.01 and 0.1 as suggestively enriched. We then used the RamiGO R package[45] for the visualization of significant GO terms within the appropriate GO tree.

RESULTS

Here, we followed a bioinformatics-based approach as summarized in Figure 1. We included the 18 SNPs that showed genome-wide significant association with CRP in the study by Dehghan et al[19] (Table 1).

Phase I: In Silico Sequencing

In this phase, we aimed to explore thoroughly the genomic area around the 18 gSNPs to identify nearby nsSNPs as potentially functional variants. We used 1000 Genomes Project data as the most detailed catalogue of human genetic variation[23]. The mini-genome of 36 Mb contains 167 003 SNPs. Of these, 3801 SNPs are in LD with the nearby gSNP at r^2>0.10, of which only 48 are exonic, including 25 nsSNPs (Table I in the Data Supplement). Of the nsSNPs, 9 map to the same gene and 16 map to other genes than the gSNPs. Please note that Tables I–III in the Data Supplement provide a thorough description of

Table 1. The 18 Genome-Wide Associated CRP SNPs Used as Primary Input to the Post-GWAS Analysis.

No. of gSNP	SNP ID	Chr	Position	Alleles	
1	rs2794520	1	159678816	C	T
2	rs4420638	19	45422946	A	G
3	rs1183910	12	121420807	G	A
4	rs4420065	1	66161461	T	C
5	rs4129267	1	154426264	C	T
6	rs1260326	2	27730940	T	C
7	rs12239046	1	247601595	T	C
8	rs6734238	2	113841030	A	G
9	rs9987289	8	9183358	A	G
10	rs10745954	12	103483094	A	G
11	rs1800961	20	43042364	C	T
12	rs340029	15	60894965	C	T
13	rs10521222	16	51158710	C	T
14	rs12037222	1	40064961	G	A
15	rs13233571	7	72971231	C	T
16	rs2847281	18	12821593	A	G
17	rs6901250	6	117114025	G	A
18	rs4705952	5	131839618	G	A

The SNPs are ordered according to the significance of their association with CRP in the meta-GWAS article. Alleles indicates ensembl reference/alternative alleles; Chr, chromosome; CRP, C-reactive protein; gSNP, GWAS SNP; GWAS, genome-wide association study; SNP, single nucleotide polymorphism; position,chromosome position build 37.

the vicinity of gSNPs by applying a liberal cutoff of r^2>0.1. These results are considered as complementary information. However, only 7 of the nsSNPs are in moderate to high LD (r^2>0.5) with the gSNPs and were used in the downstream analyses (Figure 1). The nsSNPs were then characterized for their deleterious effect on the corresponding protein function using 2 different tools, SIFT[28] and PolyPhen[29]. Interestingly, 8, 6, 4, and 10 of the nsSNPs are considered as damaging according to SIFT alone, PolyPhen alone, both SIFT and PolyPhen, or any of the 2 prediction scores, respectively (Figure 2 drawn by Circos[46]; Tables I and II in the Data Supplement).

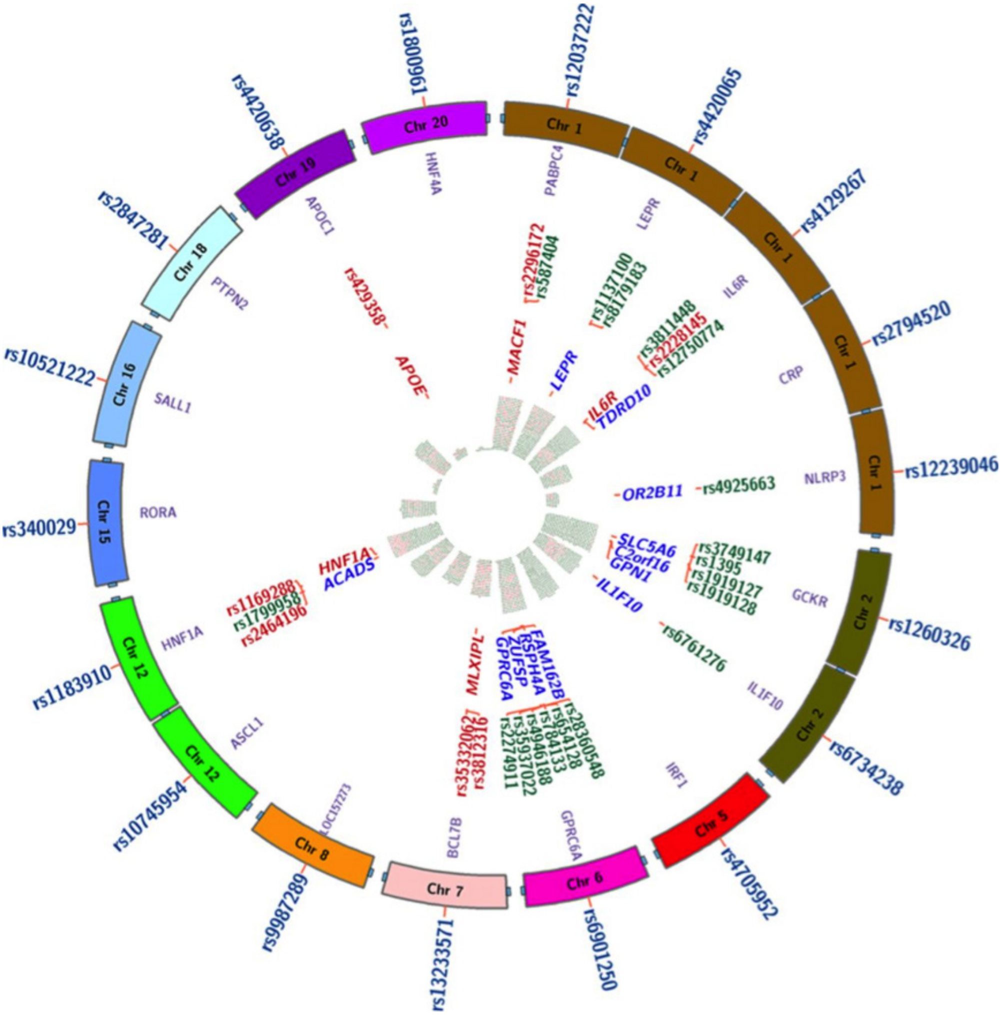


Figure 2. Results of in silico sequencing (drawn by Circos). 46 It illustrates the map of nsSNPs within the 2 Mb vicinity of 18 CRP associated SNPs. The rings from outermost to innermost represent: a) 18 CRP associated SNPs (gSNPs), b) genomic regions of 2 Mb surrounding each gSNP, c) closest genes to the gSNPs, d) 25 nsSNPs in LD with the gSNPs, e) closest genes to the nsSNPs, f) 3,801 SNPs in LD with the gSNP at r^2>0.10. The red color in rings d, e, and f indicates moderate to high LD (r^2>0.50) with the corresponding gSNP.

In silico pleiotropy analysis of all gSNPs, as well as all SNPs that are in LD with their nearby gSNP, identified several genome-wide significantly ($P<5\times10^{-8}$) associated traits or diseases other than CRP that had already been reported in previous GWAS studies as listed in the GWAS catalog[31]. By considering all gSNPs and only their highly linked variants (r2>0.80), 10 loci had effects on other traits, whereas 8 loci, including the CRP locus itself, did not show any pleiotropic effect. The locus harboring *GCKR* was the most pleiotropic region, having reported GWAS associations with a variety of metabolic-related traits. Most of the identified traits are metabolic-related traits, particularly lipid- and lipoprotein-related traits, for example, cholesterol, high-density lipoprotein, low-density lipoprotein, and triglyceride levels (Figure 3; Table III in the Data Supplement).

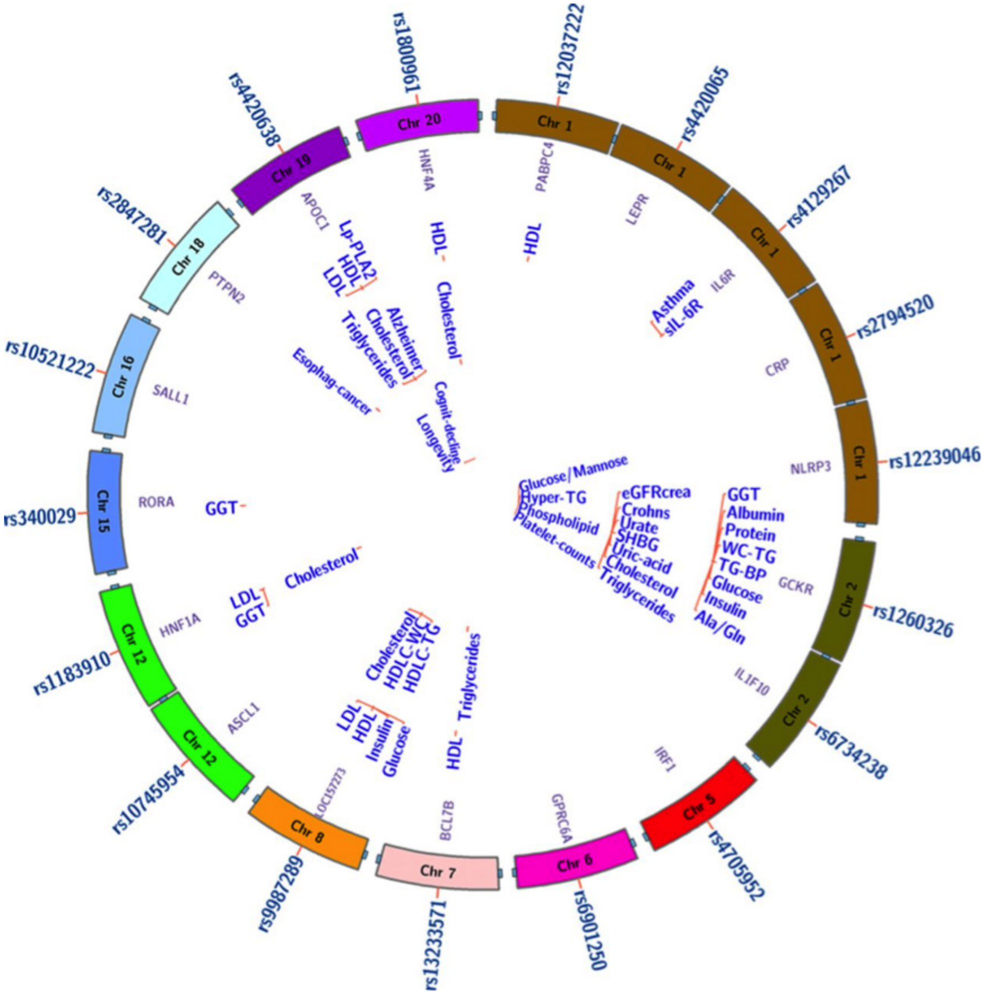


Figure 3. Results of in silico pleiotropy analysis. The three innermost rings show complex traits or diseases other than CRP, identified in previous GWAS studies to be genome-wide significantly associated with any of the gSNPs, or their highly linked variants (r2>0.80); Ala/Gln: Alanine/Glutamine; Cognit-decline: Cognitive decline; eGFRcrea: estimated glomerular filtration rate by serum creatinine; Esophag-cancer: Esophageal cancer; GGT: Gamma gluatamyl transferase; HDL: High-density lipoprotein; HDLC-TG: HDL Cholesterol-Triglycerides; HDLCWC: HDL Cholesterol-waist circumference; Hyper-TG: Hypertriglyceridemia; LDL: Lowdensity lipoprotein; Lp-PLA2: Lipoprotein-associated phospholipase A2; SHBG: Sex hormonebinding globulin; sIL-6R: Soluble Interleukin-6 receptor; TG-BP: Triglycerides-Blood Pressure; WC-TG: Waist Circumference-Triglycerides; for full trait or disease names, please see Table S3.

Phase II: eQTL Analysis

In this phase, we aimed to perform an eQTL analysis to determine whether the gSNPs affect CRP levels through regulating gene expression levels. Here we used a large data set of genome-wide expression probes in peripheral blood consisting of 5071 subjects. The eQTL analysis identified 23 expression probes that were significantly associated with 8 gSNPs at FDR<0.01. The 23 expression probes belong to 15 genes, of which 4 are the same genes and 11 are different genes from those mapping to the corresponding gSNPs. Those expression probe gene names were then used in the next step as a set of prioritized biological candidate genes (Figure 1).

Additionally, we identified the 23 SNPs that were most significantly associated with the corresponding expression probes (eSNPs; Figure 4; Table IV in the Data Supplement). eQTL analysis of the gSNPs conditional on the corresponding eSNPs revealed that for the majority of expression probes, the corresponding gSNP is not independently associated with expression levels, that is, the observed effect of gSNPs on expression probes are mostly explained by the eSNPs (Table IV in the Data Supplement).

Phase III: Network Analysis

In this phase, we generated a list of biologically prioritized candidate genes based on the findings of phases I and II as input for the construction of a functional interaction network as detailed in the methods section. Four sets of query genes were combined to create the final input list of prioritized genes for the functional interaction network analysis (Figure 1). After removing duplicate entries, the combined query gene set contained 40 genes. Two genes (*LOC157273* and *PPIEL*) could not be found in any of the available interaction resources, resulting in a final list of 38 genes (Table 2). The final composite association network contained those 38 query genes, as well as the output gene set, that is, the 100 genes connected to the query gene set. Altogether these were connected with 2225 associations, also known as edges (Figure I and Table V in the Data Supplement). All the genes in the composite network were then used for functional enrichment analysis against GO terms[47], which revealed 93 significantly (FDR<0.01) and 79 suggestively (0.1<FDR<0.01) enriched terms (Table VI in the Data Supplement). The majority of enriched terms can be broadly categorized into 2 major groups: (1) terms related to immunologic processes, cytokines, and especially interferons and (2) terms related to lipids and lipoprotein metabolism.

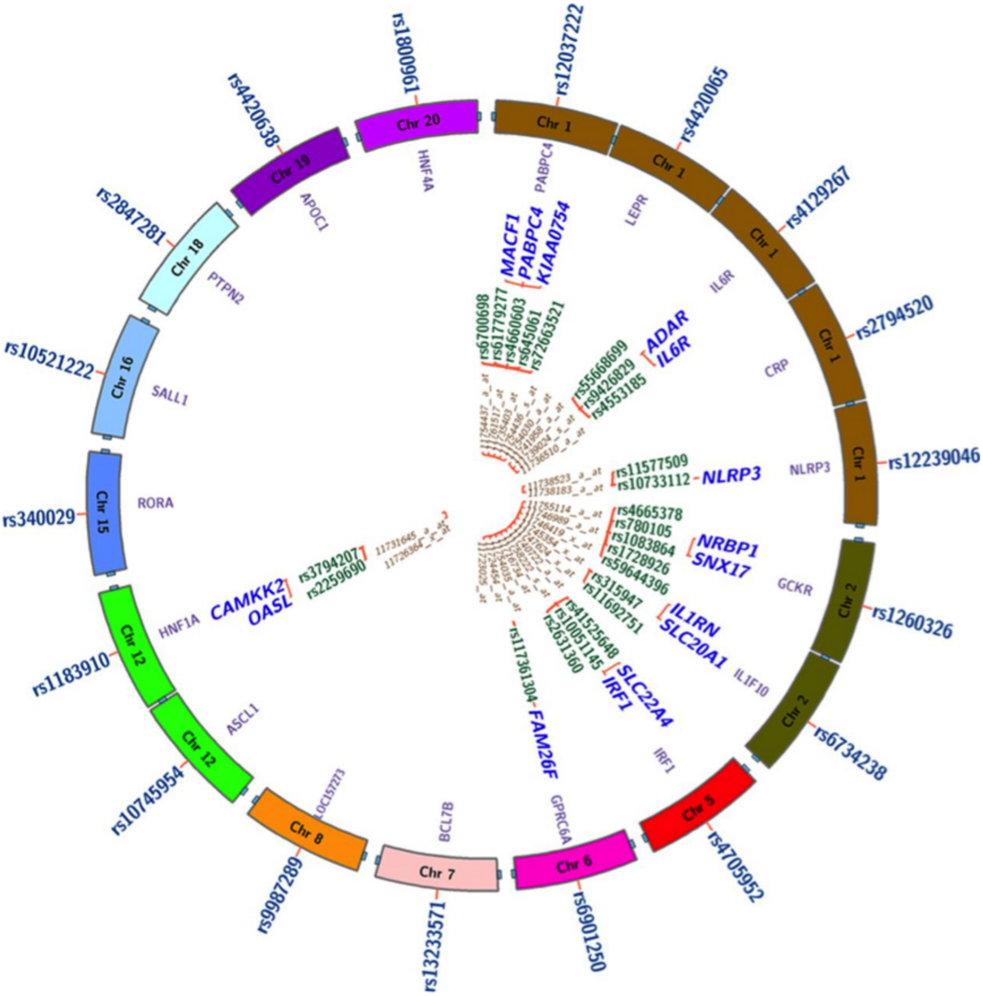


Figure 4. Results of eQTL analysis. The three innermost rings represent: d) gene names of significantly associated expression probes, e) the most significantly associated eQTL SNPs (eSNPs) for the corresponding expression probes, f) expression probes significantly associated with gSNPs.

Thirty-three of the 93 significantly enriched terms belong to the first category, of which 7 have an FDR<5×10^{-15}: cytokine-mediated signaling pathway (GO:0019221, FDR=9.47×10^{-37}), type I interferon-mediated signaling pathway (GO:0060337, FDR=1.05×10^{-34}), cellular response to type I interferon (GO:0071357, FDR=1.05×10^{-34}), response to type I interferon (GO:0034340, FDR=1.22×10^{-34}), interferon-γ–mediated signaling pathway (GO:0060333, FDR=5.78×10^{-16}), response to interferon-γ (GO:0034341, FDR=5.78×10^{-16}), cellular response to interferon-γ (GO:0071346, FDR=1.69×10^{-15}). Figure 5 visualizes these 7 terms within their corresponding GO tree. Ten out of 33 significantly enriched terms of this first category are specifically related to interferons (Table VI in the Data Supplement). Forty-three of 93 significantly enriched terms belong to the second category, that is, they are all related to the metabolism of fatty acids (eg, GO:0042304: regulation of fatty acid biosynthetic process, FDR=1.01×10^{-4}), triglycerides (eg, GO:0070328: triglyceride homeostasis, FDR=7.93×10^{-5}), cholesterol (eg, GO:0042632: cholesterol homeostasis, FDR=2.71×10^{-4}), and especially lipoproteins (eg, GO:0034361: very-low-density lipoprotein particle, FDR=3.45×10^{-5}; Table VI in the Data Supplement).

DISCUSSION

In the present study, we performed a post-GWAS analysis of 18 genome-wide significantly associated CRP SNPs. This strategy yielded new information on biological processes involved in CRP metabolism.

Here we shed light on the genomic context of the vicinity of gSNPs in 2 steps. We first investigated the nearby genomic region to identify all linked variants, with emphasis on nsSNPs as potentially functional variants. A strength of this approach is the use of r^2 as a metric of LD rather than predefined physical distance. Although nsSNPs have a high likelihood to be functional, they may constitute only a small fraction of the mechanisms involved. Therefore, we included all SNP types into the analyses. In the second step, that is, the eQTL analysis, we identified any nearby gene whose expression level is associated with its corresponding gSNP. Here we used one of the largest single data sets of genome-wide expression probes in peripheral blood currently available worldwide of >5000 samples, which was analyzed by a stringent statistical approach. The 2 steps identified several relevant genes that were jointly used as the input to the next step, that is, the functional network analysis. The strength of this approach is including the genes from the eQTL analysis in the functional network analysis, as we think these genes are at least as important as those genes to which gSNPs or their linked variants map. This approach has added value to stand-alone eQTL results as they are translated to biological insights in a broader context through integration to other data domains.

Table 2. Biologically Prioritized Candidate Gene Set Used as the Input Query to the Network Analysis.

No. of gSNP	Gene Name	Ensembl Gene ID	Query Gene Set
1	*CRP*	ENSG00000132693	i; iii
2	*APOC1*	ENSG00000130208	i; iii
2	*APOE*	ENSG00000130203	ii
2	*APOC1P1*	ENSG00000214855	iii
3	*HNF1A*	ENSG00000135100	i; ii; iii
3	*CAMKK2*	ENSG00000110931	iv
3	*OASL*	ENSG00000135114	iv
4	*LEPR*	ENSG00000116678	i; iii
5	*IL6R*	ENSG00000160712	i; ii; iii; iv
5	*ADAR*	ENSG00000160710	iv
6	*GCKR*	ENSG00000084734	i; iii
6	*NRBP1*	ENSG00000115216	iv
6	*SNX17*	ENSG00000115234	iv
7	*NLRP3*	ENSG00000162711	i; iii; iv
8	*IL1F10*	ENSG00000136697	i; iii
8	*IL1RN*	ENSG00000136689	iii; iv
8	*SLC20A1*	ENSG00000144136	iv
9	*LOC157273*	ENSG00000254235	i; iii
10	*ASCL1*	ENSG00000139352	i; iii
10	*C12orf42*	ENSG00000179088	iii
11	*HNF4A*	ENSG00000101076	i
12	*RORA*	ENSG00000069667	i; iii
13	*SALL1*	ENSG00000103449	i; iii
14	*PABPC4*	ENSG00000090621	i; iii; iv
14	*MACF1*	ENSG00000127603	ii; iv
14	*HEYL*	ENSG00000163909	iii
14	*PPIEL*	ENSG00000243970	iii
14	*BMP8A*	ENSG00000183682	iii
14	*KIAA0754*	ENSG00000255103	iv
15	*BCL7B*	ENSG00000106635	i; iii
15	*MLXIPL*	ENSG00000009950	ii; iii
15	*BAZ1B*	ENSG00000009954	iii
15	*TBL2*	ENSG00000106638	iii
16	*PTPN2*	ENSG00000175354	i; iii
17	*GPRC6A*	ENSG00000173612	i; iii
17	*RFX6*	ENSG00000185002	iii
17	*FAM162B*	ENSG00000183807	iii
17	*FAM26F*	ENSG00000188820	iv
18	*IRF1*	ENSG00000125347	i; iv
18	*SLC22A4*	ENSG00000197208	iv

The query gene set includes the following: (i) closest genes to the 18 gSNPs, (ii) closest genes to the nsSNPs in high LD (r^2>0.50) with the corresponding gSNP, (iii) closest genes to other types of SNPs in very high LD (r^2>0.80) with the corresponding gSNP, and (iv) expression probe gene names significantly associated with gSNPs (FDR<0.01) based on the eQTL analysis. The combined query gene set contained 40 genes, of which, 2 genes, ***LOC157273*** and ***PPIEL***, could not be found in any of the interaction resources. The order of genes follows the order of gSNPs in Table 1. eQTL indicates expression quantitative trait loci;FDR, false discovery rate; GWAS, genome-wide association study; LD, linkage disequilibrium; gSNP, GWAS SNPs; nsSNP, nonsynonymous SNPs; and SNP,single nucleotide polymorphism.

In the next step, we constructed a functional association interaction network followed by functional enrichment analysis against GO terms. Such an interaction network is considered to represent cofunctionality of the connected genes[41]. The large data set of functional association data that is used contains not only coexpression data, but also physical interaction, genetic interaction, shared protein domains, colocalization, and predicted association networks. As a result, the constructed interaction network is a composite based on these different data sources[41]. As described in the methods section, the functional enrichment analysis is performed against GO terms after excluding those annotations with computer-generated inferred from RCA and IEA evidence codes. Thus, about half of the GO annotations are disregarded to avoid circularity and to obtain more robust results (http://www.geneontology.org/GO.evidence.shtml)[44].

Our post-GWAS analysis of CRP GWAS SNPs eventually yielded a range of enriched biological processes after several intermediate steps. Some processes like acute-phase response or acute inflammatory response with significant FDR values are expected and appropriate terms for CRP providing confidence in our results. Interestingly, about one third of the significantly enriched terms were related to immunologic processes, cytokines, and interferons. Even more interesting, 10 of the significantly enriched terms, including 6 of the top most significant ones, are those pointing to the biology of interferons. In particular, type I interferon associated biological processes are highlighted with 3 significant enriched terms with $FDR<1\times10^{-30}$.

The link between interferons and CRP has not been well established, probably because the measurement of interferons is complicated by their short half-lives. Although few studies have addressed the direct link between CRP and interferons and although this link has not been appreciated as a potential mechanism underlying the biology of CRP, our finding is in fact amply supported by those few in vitro and clinical observations. An in vitro observation by Enocsson et al showed that interferon-α, the main representative of the type I interferon family, inhibits CRP secretion in a dose-dependent fashion mediated by the type I interferon receptor[48]. Furthermore, although CRP levels are highly associated with most inflammatory states, as CRP level is a well-known metric for the detection and evaluation of many inflammatory diseases[10], elevated CRP levels correlate poorly with those inflammatory conditions that are characterized by high levels

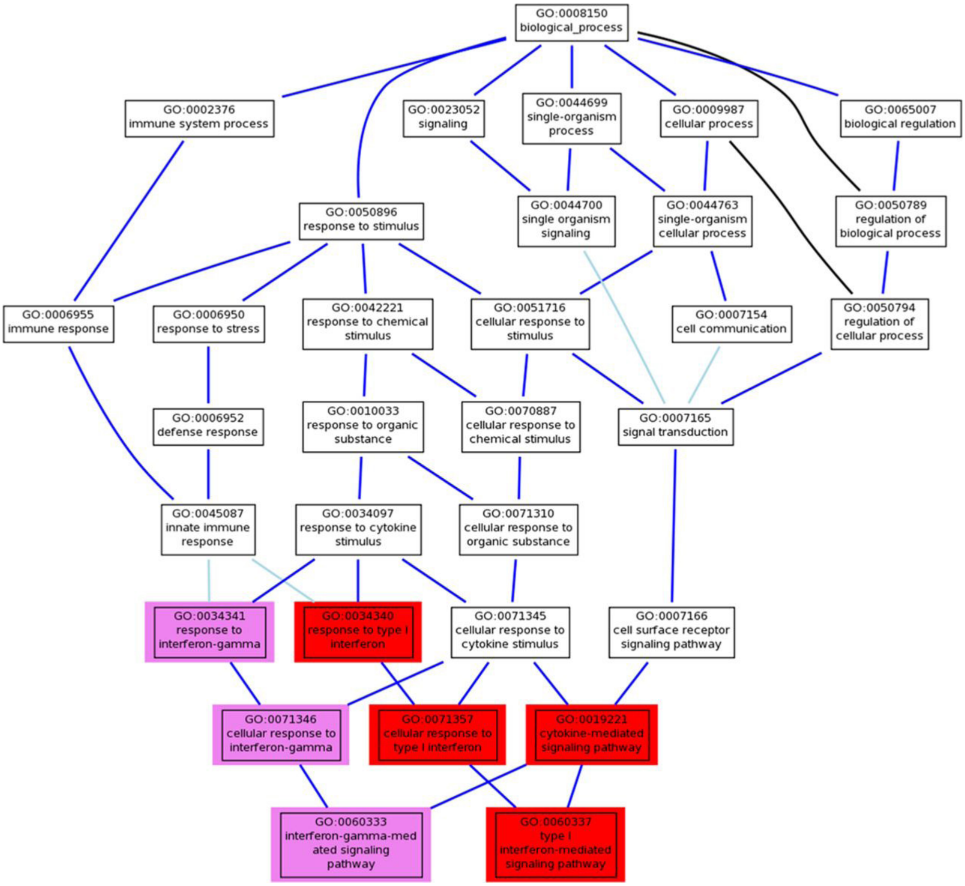


Figure 5. The most significantly enriched GO terms with FDR$<5\times10^{-15}$. They are visualized as highlighted boxes within their corresponding GO tree, as red for those with FDR$<1\times10^{-30}$ and purple for those with $5\times10^{-15}<$FDR$<1\times10^{-30}$. The relations between the boxes have standard colors: black (regulates), blue (is_a) or light blue (part_of) (http://www.geneontology.org/GO.ontology-ext.relations.shtml).

of interferon-α, such as systemic lupus and viral infections[49–54]. This observation is in line with the abovementioned in vitro observation that increased levels of interferon-α suppress CRP levels[48]. Likewise, there are yet unexplained phenomena in lupus patients as there is a 10- to 50-fold increased risk of myocardial infarction[55,56], whereas there is no association between cardiovascular disease and CRP levels in these patients[53]. This lack of an association is unexpected because CRP is an established risk factor for coronary heart disease[5,6]. Moreover, in lupus patients, lack of correlation between interleukin-6, the main stimulant of CRP secretion, and CRP has been reported[57]. These related observations may be explained by the fact that lupus patients are known to have a high level of interferon-α and that interferon-α is an inhibitor of CRP secretion. Another line of evidence comes from infectious diseases. In viral infections, in contrast to bacterial infections, there is generally a mild, poorly correlated increase of CRP level, making CRP a widely used diagnostic tool in distinguishing viral from bacterial infections[54]. This can be explained by the notion that patients with viral infections have high levels of interferon-α 49 and by an inverse relation between interferon-α and CRP levels[48]. Finally, although the analysis had started with CRP gSNPs, it interestingly returned 4 significantly enriched terms specifically related to defense responses to viruses (Table VI in the Data Supplement). Considering the blunted response of CRP levels to viral infections[54], this unexpected finding once again suggests an important role of interferon-α in CRP metabolism.

The in silico pleiotropy analysis revealed several pleiotropic effects between CRP gSNPs and other metabolic traits, particularly lipid- and lipoprotein-related traits. These results show strong concordance with those from our functional network analysis, as about half of the significantly enriched GO terms point to biological processes related to lipids and lipoproteins metabolism. These findings are also fully in line with existing knowledge of overlap between the biology of CRP and lipids with metabolism of both CRP and lipids related to the liver. Further, CRP levels are significantly associated with weight, waist-circumference, body mass index, cholesterol, triglycerides, low-density lipoprotein (weakly) and negatively associated with high-density lipoprotein concentrations[6,58–61]. Both CRP and lipids are well-known risk factors for coronary heart disease[61]. Thus, our results show extensive genetic overlap between CRP and lipid metabolism, although the exact mechanisms underlying these significant associations remain to be elucidated.

In early 2010, Dickson et al suggested that observed GWAS associations between a common SNP and trait of interest can be explained by multiple rare variants at the locus in LD with that SNP, so-called synthetic associations[62]. However, there are several lines of evidence indicating that GWAS associations are rarely caused by synthetic associations with rare variants[63–65]. Later on, Visscher and colleagues state that instead the combined evidence supports a highly polygenic model of disease susceptibility which is built on causal variants across the entire range of the allele-frequencies[66]. Hence, our approach of including all gSNPs, as well as their linked SNPs and eQTL results, is more consistent with the polygenic model than with the synthetic association model.

Despite using one of the largest single data sets of genome-wide expression probes for eQTL analysis, it contained only blood expression probes. This limitation may have affected the list of associated genes. A similar approach but using a large data set of tissue-specific expression data, particularly liver cells of healthy individuals, may better reveal the associated gene expressions. However, to the best of our knowledge, such a homogenous large data set of liver cells from healthy individuals does not exist yet. Furthermore, if there is cryptic relatedness among our subjects, it is possible that our eQTL results might be slightly biased. However, our population is relatively outbred and known relationships among subjects were taken into account in the analysis. Under these circumstances, Voight and Pritchard suggest that the bias is expected to be negligible[67]. Our functional enrichment analysis was done using GO terms; one may suggests a more extended approach by including other annotation sources like KEGG and Reactome pathways. However, as these resources only contain a limited number of pathways, it is unlikely this would have affected our main conclusions.

Finally, the results of this in silico study need to be followed up by further in vitro, in vivo, and epidemiological studies. The association of interferon-α with coronary heart disease and other CRP-associated traits or diseases, as well as the association of CRP gSNPs or CRP genetic risk scores with clinical conditions like systemic lupus, are yet to be investigated. These results also highlight the need and potential for a GWAS on serum levels of interferon-α. Finally, although those CRP gSNPs are based on a large meta-GWAS, including >80 000 subjects, the explained variance in CRP level by all those 18 gSNPs is only around 5%[19]. To further unravel the underlying genetic mechanisms controlling CRP levels, a larger meta-GWAS on CRP is needed to find additional common variants, whereas other approaches, such as meta-analyses of exome chip data, will be needed to find variants of lower frequency affecting serum levels of CRP.

In summary, in this in silico study, we followed a bioinformatics-based approach aiming to translate CRP GWAS signals into biological insights. Our post-GWAS analysis of CRP GWAS SNPs reemphasizes the previously known overlap between the biology of CRP and lipids. Additionally, it suggests an important role for interferons in the metabolism of CRP.

Accession numbers:
Gene expression and genotype data used for this study will be available at dbGaP, accession number phs000486.v1.p1 (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000486.v1.p1).

Sources of funding

The Netherlands Study of Depression and Anxiety (NESDA) and the Netherlands Twin Register (NTR) were funded by the Netherlands Organization for Scientific Research (MagW/ZonMW grants 904-61-090, 985-10-002, 904-61-193, 480-04-004, 400-05-717, 912-100-20; Spinozapremie 56-464-14192; Geestkracht program grant 10-000-1002); the Center for Medical Systems Biology (NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure, VU University's Institutes for Health and Care Research and Neuroscience Campus Amsterdam, NBIC/BioAssist/RK (2008.024); the European Science Foundation (EU/QLRT-2001-01254); the European Community's Seventh Framework Program (FP7/2007–2013); ENGAGE (HEALTH-F4-2007–201413); and the European Science Council (ERC, 230374). Gene-expression data were funded by the US National Institute of Mental Health (RC2 MH089951) as part of the American Recovery and Reinvestment Act of 2009. NESDA and NTR were further supported by funding from the Neuroscience Campus Amsterdam and through a NWO-VICI grant (number 91811602) and funding from ZonMW Middelgroot (911-09-032).

Supplementary material

Supplementary Material is available at Circulation Cardiovascular Genetics online.

REFERENCES

1. Nordestgaard, B. G. Does elevated C-reactive protein cause human atherothrombosis? Novel insights from genetics, intervention trials, and elsewhere. Curr. Opin. Lipidol. 20, 393–401 (2009).
2. Allin, K. H., Bojesen, S. E. & Nordestgaard, B. G. Baseline C-reactive protein is associated with incident cancer and survival in patients with cancer. J. Clin. Oncol. 27, 2217–2224 (2009).
3. Dehghan, A. et al. Genetic variation, C-reactive protein levels, and incidence of diabetes. Diabetes 56, 872–878 (2007).
4. Sesso, H. D. et al. C-reactive protein and the risk of developing hypertension. JAMA 290, 2945–2951 (2003).
5. Danesh, J. et al. C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. N. Engl. J. Med. 350, 1387–1397 (2004).
6. Emerging Risk Factors Collaboration et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet 375, 132–140 (2010).
7. De Berardis, D. et al. Evaluation of C-reactive protein and total serum cholesterol in adult patients with bipolar disorder. Int J Immunopathol Pharmacol 21, 319–324 (2008).
8. Harris, T. B. et al. Associations of elevated interleukin-6 and C-reactive protein levels with mortality in the elderly. Am. J. Med. 106, 506–512 (1999).
9. Tremblay, J. Genetic determinants of C-reactive protein levels in metabolic syndrome: a role for the adrenergic system? J. Hypertens. 25, 281–283 (2007).
10. Ummarino, D. & Zeng, L. Is C reactive protein expression affected by local microenvironment? Heart 99, 514–515 (2013).
11. Danik, J. S. & Ridker, P. M. Genetic determinants of C-reactive protein. Curr Atheroscler Rep 9, 195–203 (2007).
12. Elliott, P. et al. Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. JAMA 302, 37–48 (2009).
13. Kathiresan, S. et al. Contribution of clinical correlates and 13 C-reactive protein gene polymorphisms to interindividual variability in serum C-reactive protein level. Circulation 113, 1415–1423 (2006).
14. Saunders, C. L. & Gulliford, M. C. Heritabilities and shared environmental effects were estimated from household clustering in national health survey data. J Clin Epidemiol 59, 1191–1198 (2006).
15. Rahman, I. et al. Genetic dominance influences blood biomarker levels in a sample of 12,000 Swedish elderly twins. Twin Res Hum Genet 12, 286–294 (2009).
16. Su, S. et al. Common genetic contributions to depressive symptoms and inflammatory markers in middle-aged men: the Twins Heart Study. Psychosom Med 71, 152–158 (2009).
17. Neijts, M. et al. Genetic architecture of the pro-inflammatory state in an extended twin-family design. Twin Res Hum Genet 16, 931–940 (2013).
18. Ridker, P. M. et al. Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and *GCKR* associate with plasma C-reactive protein: the Women's Genome Health Study. Am. J. Hum. Genet. 82, 1185–1192 (2008).
19. Dehghan, A. et al. Meta-analysis of genome-

wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123, 731–738 (2011).

20. Wang, X., Prins, B. P., Sõber, S., Laan, M. & Snieder, H. Beyond genome-wide association studies: new strategies for identifying genetic determinants of hypertension. Curr. Hypertens. Rep. 13, 442–451 (2011).
21. Freedman, M. L. et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 43, 513–518 (2011).
22. Franceschini, N. et al. Discovery and fine mapping of serum protein loci through transethnic meta-analysis. Am. J. Hum. Genet. 91, 744–753 (2012).
23. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).
24. Meyer, L. R. et al. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 41, D64–69 (2013).
25. Danecek, P. et al. The variant call format and VCFtools. Bioinformatics 27, 2156–2158 (2011).
26. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27, 718–719 (2011).
27. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164 (2010).
28. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4, 1073–1081 (2009).
29. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
30. Flicek, P. et al. Ensembl 2012. Nucleic Acids Res. 40, D84–90 (2012).
31. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. 106, 9362–9367 (2009).
32. Jansen, R. et al. Sex differences in the human peripheral blood transcriptome. BMC Genomics 15, 33 (2014).
33. Wright, F. A. et al. Heritability and genomics of gene expression in peripheral blood. Nat. Genet. 46, 430–437 (2014).
34. Penninx, B. W. J. H. et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatr Res 17, 121–140 (2008).
35. Boomsma, D. I. et al. Netherlands Twin Register: from twins to twin families. Twin Res Hum Genet 9, 849–857 (2006).
36. Boomsma, D. I. et al. Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. Eur. J. Hum. Genet. 16, 335–342 (2008).
37. Abdellaoui, A. et al. Population structure, migration, and diversifying selection in the Netherlands. Eur. J. Hum. Genet. 21, 1277–1285 (2013).
38. Fehrmann, R. S. N. et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 7, e1002197 (2011).
39. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. 45, 1238–1243 (2013).

40. Visscher, P. M., Benyamin, B. & White, I. The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. Twin Res 7, 670–674 (2004).

41. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 9 Suppl 1, S4 (2008).

42. Saito, R. et al. A travel guide to Cytoscape plugins. Nat. Methods 9, 1069–1076 (2012).

43. Montojo, J. et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 26, 2927–2928 (2010).

44. Mostafavi, S. & Morris, Q. Combining many interaction networks to predict gene function and analyze gene lists. Proteomics 12, 1687–1696 (2012).

45. Schröder, M. S., Gusenleitner, D., Quackenbush, J., Culhane, A. C. & Haibe-Kains, B. RamiGO: an R/Bioconductor package providing an AmiGO visualize interface. Bioinformatics 29, 666–668 (2013).

46. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645 (2009).

47. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000).

48. Enocsson, H. et al. Interferon-alpha mediates suppression of C-reactive protein: explanation for muted C-reactive protein response in lupus flares? Arthritis Rheum. 60, 3755–3760 (2009).

49. Theofilopoulos, A. N., Baccala, R., Beutler, B. & Kono, D. H. Type I interferons (alpha/beta) in immunity and autoimmunity. Annu. Rev. Immunol. 23, 307–336 (2005).

50. Lech, M., Rommele, C. & Anders, H.-J. Pentraxins in nephrology: C-reactive protein, serum amyloid P and pentraxin-3. Nephrol. Dial. Transplant. 28, 803–811 (2013).

51. Becker, G. J., Waldburger, M., Hughes, G. R. & Pepys, M. B. Value of serum C-reactive protein measurement in the investigation of fever in systemic lupus erythematosus. Ann. Rheum. Dis. 39, 50–52 (1980).

52. Honig, S., Gorevic, P. & Weissmann, G. C-reactive protein in systemic lupus erythematosus. Arthritis Rheum. 20, 1065–1070 (1977).

53. Nikpour, M., Gladman, D. D., Ibañez, D. & Urowitz, M. B. Variability and correlates of high sensitivity C-reactive protein in systemic lupus erythematosus. Lupus 18, 966–973 (2009).

54. Sasaki, K., Fujita, I., Hamasaki, Y. & Miyazaki, S. Differentiating between bacterial and viral infection by measuring both C-reactive protein and 2′-5′-oligoadenylate synthetase as inflammatory markers. J. Infect. Chemother. 8, 76–80 (2002).

55. Manzi, S. et al. Age-specific incidence rates of myocardial infarction and angina in women with systemic lupus erythematosus: comparison with the Framingham Study. Am. J. Epidemiol. 145, 408–415 (1997).

56. Esdaile, J. M. et al. Traditional Framingham risk factors fail to fully account for accelerated atherosclerosis in systemic lupus erythematosus. Arthritis Rheum. 44, 2331–2337 (2001).

57. Gabay, C. et al. Absence of correlation between interleukin 6 and C-reactive protein blood levels in systemic lupus erythematosus compared with rheumatoid arthritis. J. Rheumatol. 20, 815–821 (1993).

58. Mendall, M. A., Patel, P., Ballam, L., Strachan, D. & Northfield, T. C. C reactive protein and its relation to cardiovascular risk factors: a population based cross sectional study. BMJ 312, 1061–1065 (1996).

59. Kraja, A. T. et al. Do inflammation and procoagulation biomarkers contribute to the metabolic syndrome cluster? Nutr Metab (Lond) 4, 28 (2007).

60. Sakkinen, P. A., Wahl, P., Cushman, M., Lewis, M. R. & Tracy, R. P. Clustering of procoagulation, inflammation, and fibrinolysis variables with metabolic factors in insulin resistance syndrome. Am. J. Epidemiol. 152, 897–907 (2000).

61. Ridker, P. M., Rifai, N., Rose, L., Buring, J. E. & Cook, N. R. Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. N. Engl. J. Med. 347, 1557–1565 (2002).

62. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. PLoS Biol. 8, e1000294 (2010).

63. Anderson, C. A., Soranzo, N., Zeggini, E. & Barrett, J. C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol. 9, e1000580 (2011).

64. Wray, N. R., Purcell, S. M. & Visscher, P. M. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. 9, e1000579 (2011).

65. Orozco, G., Barrett, J. C. & Zeggini, E. Synthetic associations in the context of genome-wide association scan signals. Hum. Mol. Genet. 19, R137–144 (2010).

66. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. Am J Hum Genet 90, 7–24 (2012).

67. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1, e32 (2005).

SECTION II:

INTEGRATIVE POST-GWAS ANALYSES AND SYSTEMS GENETICS

7

Investigating the causal relationship of C-reactive protein with 32 complex somatic and psychiatric outcomes: A large scale cross-consortia Mendelian randomization study.

PLOS Medicine. 2016 Jun 21;13(6):e1001976

Bram. P. Prins, Ali Abbasi[#], Anson Wong[#], Ahmad Vaez[#], Ilja Nolte, Nora Franceschini, Philip E. Stuart, Javier Guterriez-Achury, Vanisha Mistry, Jonathan P. Bradfield, Ana M. Valdes, Jose Bras, Aleksey Shatunov, PAGE consortium, International Stroke Genetics Consortium, Systemic Sclerosis consortium, Treat OA consortium, DIAGRAM consortium, CARDIoGRAMplusC4D consortium, ALS consortium, International Parkinson's Disease Genomics Consortium, Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium, CKDGen consortium, the GERAD1 consortium, ICBP consortium, Schizophrenia Working Group of the Psychiatric Genomics consortium, Inflammation working group of CHARGE consortium, Chen Lu, Buhm Han, Soumya Raychaudhuri, Steve Bevan, Maureen D. Mayes, Lam C. Tsoi, Evangelos Evangelou, Rajan P. Nair, Struan F.A. Grant, Constantin Polychronakos, Timothy R.D. Radstake, David A. van Heel, Melanie L. Dunstan, Nicholas W. Wood, Ammar Al-Chalabi, Abbas Dehghan, Hakon Hakonarson, Hugh S. Markus, James T. Elder, Jo Knight, Dan E. Arking. Timothy D. Spector, Bobby P.C. Koeleman, Cornelia M. van Duijn, Javier Martin, Andrew P. Morris, Rinse K. Weersma, Cisca Wijmenga, Patricia B. Munroe, John R.B. Perry, Jennie G. Pouget, Yalda Jamshidi, Harold Snieder, and Behrooz Z. Alizadeh.

[#] co-authors that contributed equally.

ABSTRACT

Background

C-reactive protein (CRP) is associated with immune, cardiometabolic and psychiatric traits and diseases. Yet it is inconclusive whether these associations are causal.

Methods and Findings

We performed Mendelian randomization (MR) analyses using two genetic risk scores (GRS) as instrumental variables (IVs). The first consisted of four single nucleotide polymorphisms (SNPs) in the CRP gene (GRS_{CRP}), and the second of eighteen SNPs that were significantly associated with CRP levels in the largest genome-wide association study (GWAS; by Dehghan A. et al 2011) to date (GRS_{GWAS}). To optimize power we used summary statistics from GWAS consortia and tested association of these two GRSs with 32 complex somatic and psychiatric outcomes comprising up to 123,865 participants per outcome from populations of European ancestry. We performed heterogeneity tests to disentangle pleotropic effect of IVs. A Bonferroni corrected significance level of less than 0.0016 was considered statistically significant. An observed P value equal or less than 0.05 as nominal significant evidence for a potential causal association but yet to be confirmed.

The strengths (F-statistics) of IVs were between 31.92-3761.29 and 82.32-9403.21 for GRS_{CRP} and GRS_{GWAS}, respectively. CRP GRS_{GWAS} showed a statistically significant protective relationship of a 10% genetically elevated CRP levels with the risk of schizophrenia (OR 0.86 [95% CI 0.79-0.94];$P<0.001$). We validated this finding with individual-level genotype data from the schizophrenia GWAS (0.96 [0.94-0.98];$P<1.72x10^{-6}$). Further, we found that standardized CRP polygenic risk scores (CRP_{PRS}) at P value thresholds of 1x10-4, 0.001, 0.01, 0.05, and 0.1 using individual levels data, also showed a protective effect (OR<1.00) against schizophrenia; when the first CRP_{PRS} (built of SNPs with a P value$<1x10^{-4}$) showed a statistically significant ($P<2.45x10^{-4}$) protective effect with an OR of 0.97 [0.95-0.99]. The CRP GRS_{GWAS} showed that a 10% increase in genetically-determined CRP levels showed a significant association with coronary artery disease (OR 0.88 [0.84-0.94];$P<2.4x10^{-5}$) and had a nominal associations with risk of IBD (OR 0.85 [95%CI 0.74–0.98];P<0.03), Crohn's disease (0.81[0.70-0.94];P<0.005), psoriatic arthritis (1.36[1.00-1.84];P<0.049), knee osteoarthritis (1.17[1.01-1.36];P<0.04), bipolar disorder (1.21[1.05-1.40];P<0.007), and with an increase of 0.72 (0.11-1.34;P<0.02) mmHg in systolic blood pressure, 0.45 (0.06-0.84;P<0.02) mmHg in diastolic blood pressure, 0.01 ml/min/1.73m2 (0.003-0.02;P<0.005) in estimated creatinine glomerular filtration rate, 0.01 g/dl (0.0004-0.02;P<0.04) in albumin, and 0.03 g/dl (0.008-0.05;P<0.009) in serum protein levels. However, after adjustment for heterogeneity, no GRS showed a significant effect (at p<0.0016) on any of these outcomes, including CAD, nor on the other 20 complex outcomes studied. Our study has two potential limitations; firstly the limited variance explained by our genetic instruments modelling CRP levels in blood, and unobserved bias introduced by the use of summary statistics in our MR analyses.

Conclusions

Genetically elevated CRP levels showed a significant potentially protective causal relationship with risk of schizophrenia. We observed nominal, yet to be confirmed, evidence for a causal relationship of elevated CRP levels with psoriatic osteoarthritis, rheumatoid arthritis, knee osteoarthritis, SBP, DBP, serum albumin, and bipolar disorder. We cannot verify any causal effect of CRP on other common somatic and neuropsychiatric outcomes investigated in the present study. This implies that interventions lowering CRP levels are unlikely to result in decreased risk for the majority of common complex outcomes.

INTRODUCTION

Emerging evidence suggests that the persistent dysregulation of the inflammatory response is linked to a plethora of complex somatic and neuropsychiatric disorders[1–18]. Epidemiological studies have shown that C-reactive protein (CRP), a well-studied biomarker of inflammation, is associated with and exhibited a reliable predictive value for cardiovascular disease[19,20], type 2 diabetes[21], and immunity-related disorders such as inflammatory bowel disease (IBD)[22], rheumatoid arthritis[23] and all-cause mortality[20,24]. Nevertheless, the evidence for a causal involvement of CRP from traditional experimental or observational studies remains controversial[25,26], fuelling the debate surrounding whether CRP contributes to the chain of causality in disease mechanisms[27]. The use of genetically informed instrumental variables (IVs) termed Mendelian randomization is a complementary approach to epidemiological observations and allows investigating whether the effect of an exposure (i.e. CRP levels) on observed outcome phenotypes is likely to be causal[28].

Recent large-scale Mendelian randomization studies, focussing mainly on cardiovascular disease and metabolic traits, failed to show a causal association between CRP and these outcomes (S1 Table). This has led to the notion that elevated CRP levels do not causally contribute to these traits and disorders. However, these studies have used either a single CRP-associated single nucleoid polymorphism (SNP), or a very limited set of CRP-associated SNPs (S1 Table). Common SNPs serving as proxies for CRP levels represent only a small effect on CRP levels *per se* and thus require a large enough sample size to detect causal effects on the outcome. Moreover, most studies have generally included a limited range of common, complex diseases, often not more than two or three outcomes, or they have been performed in a single or small population yielding inadequate study power (S1 Table). In other words, existing evidence for a causal relationship between CRP and a broad range of common traits or diseases remains inconclusive. This is mostly due to the lack of well-powered Mendelian randomization studies that use optimally informative genetic IVs for CRP. Here, we sought to comprehensively examine the hypothesis that genetically determined CRP levels directly contribute to common somatic and psychiatric outcomes. To optimize IV power, we applied a Mendelian randomization approach using summary statistics from large-scale genome-wide association study (GWAS) consortia of 32 somatic and psychiatric phenotypes for the four variants representing 98% of the common variation in the *CRP* gene, and the largest known set of independent SNPs known to be associated with CRP. We further aimed to confirm the identified association between CRP and schizophrenia using CRP polygenic risk score (CRP_{PRS}) from individual level genotype data of the largest consortium of schizophrenia to-date. We performed an *Insilco* pathway analysis (see Discussion) to speculate the possible mechanism underlying the observed associations with schizophrenia.

METHODS

Study design and rationale

The present Mendelian randomization study consists of two key components: first, we used established variants associated with CRP levels, and combined them to build two genetic risk scores (GRS) for CRP: The first one consisted of only four SNPs in the *CRP* gene (GRS_{CRP}) selected from the largest recent Mendelian randomization study of CRP[29], and the second consisted of 18 SNPs that were associated with CRP levels at genome-wide significance in the largest GWAS for CRP to date (GRS_{GWAS})[30]. Second, we obtained summary association statistics from GWAS consortia for a panel of 32 common somatic and psychiatric outcomes (Table 1). The corresponding authors selected the studies, and contacted each consortium with a standardized request for study data, including the name of study or consortium, number of cases and controls, number of available CRP SNPs for GRS_{CRP} and GRS_{GWAS}, and the estimated effects for each SNP (or its proxy) on outcome, i.e. per allele regression coefficient with standard errors or odds ratio and corresponding 95% confidence interval. Data were available for 32 different outcomes in five broad disease classes, (i.e. auto-immune-inflammatory, cardiovascular, metabolic, neuro-degenerative and psychiatric), including at least 1,566 up to 184,305 participants per outcome from populations of European ancestry (Table 1). These outcomes were selected based on the following two inclusion criteria: (i) having been associated with CRP levels in epidemiological studies and (ii) availability of large meta-GWAS analyses for the outcome (Table 1).

Genetic instruments

Weak IVs yielding insufficient statistical power may have hampered estimation of causal effects of CRP on the outcomes in previous analyses (S1 Table). Our Mendelian randomization approach, by using GWAS data, and combining multiple independent

Table 1. Diseases and traits included in this study.

Disease / Trait Class	Abbreviation	Cases	Controls	Total	Reference
Autoimmune/Inflammatory					
Celiac Disease	CED	4533	10750	15283	31
Inflammatory Bowel Disease (all types)	IBD	13020	34774	47794	32,33
Crohn's Disease	CD	6333	15056	21389	32
Ulcerative Colitis	UC	6687	19718	26405	33
Psoriasis Vulgaris	PSV	4007	4934	8941	34,35
Psoriatic Arthritis	PSA	1946	4934	6880	34,35
Psoriasis Cutaneous	PSC	1363	3517	4880	34,35
Rheumatoid Arthritis	RA	5538	20167	25705	36
Systemic Lupus Erythematous	SLE	1311	3340	4651	37
Systemic Sclerosis	SSC	2356	5187	7543	38
Type 1 Diabetes	T1D	9934	16956	26890	39
Knee Osteoarthritis	KOA	5755	18505	24260	40
Cardiovascular					
Coronary Artery Disease	CAD	60801	123504	184305	41
Systolic Blood Pressure	SBP	–	–	69368	42
Diastolic Blood Pressure	DBP	–	–	69372	42
Ischemic Stroke (all types)	IS	3548	5972	9520	43
Ischemic Stroke (Cardioembolic)	IS (CS)	790	5972	6762	43
Ischemic Stroke (Large Vessel)	IS (LVS)	844	5972	6816	43
Ischemic Stroke (Small Vessel)	IS (SVD)	580	5972	6522	43
Metabolic					
Body Mass Index	BMI	-	-	123865	44
Type 2 Diabetes	T2D	6698	15872	22570	45
Chronic Kidney Disease	CKD	6271	68083	74354	46
eGFR for creatinine	eGFR	-	-	74354	46
Serum Albumin Levels	SA	-	-	53189	47
Serum Protein Levels	SP	-	-	25537	47
Neurodegenerative					
Amyotrophic Lateral Sclerosis	ALS	4133	8130	12663	48
Alzheimer's Disease	ALZ	4663	8357	13020	49
Parkinson's Disease	PKD	5333	12019	17352	50
Psychiatric					
Autism	AUT	90	1476	1566	51
Bipolar Disorder	BPD	7481	9250	16731	52
Major Depressive Disorder	MDD	9240	9519	18759	53
Schizophrenia	SCZ	34241	45604	79845	54

SNPs into a GRS (i.e. IV), has the potential to greatly increase power. The selected SNPs have been described elsewhere[30,55,56] and are further detailed in S2 Table, S3 Table and S4 Table. These IVs were used to test the combined effect of the associations of CRP level influencing alleles with the outcomes. Our approach was implemented in such a way that both the effects of independent SNPs in the *CRP* gene (GRS_{CRP})[55,56] (S1 Methods) and independent SNPs known to be genome-wide significantly associated with CRP levels (GRS_{GWAS})[30], as well as pleiotropic effects of SNPs could be discriminated[57]. Pleiotropy exists if CRP SNPs influence exposures (risk factors) other than CRP levels and therefore would violate one of the key Mendelian randomization assumptions.

Statistical analysis

All analyses were done using the GRS function implemented in the grs.summary module as part of the Genetics ToolboX R (version 2.15.1 for Windows; Vienna, Austria). The grs.summary module approximates the regression of an outcome onto an additive GRS, using only single SNP association summary statistics extracted from GWAS results. The method is described in more detail elsewhere[58]. In brief, we performed Mendelian randomization analyses using GRS IVs in two steps: First, we used individual *CRP* gene SNPs (i.e. IVs) associated with CRP levels[56,59] (S2 Table and S3 Table) to create a weighted GRS, named $GRS_{CRP,}$ corresponding to the joint effect of four SNPs within the *CRP* gene[55]. We extracted ω (the estimated coefficient or weight) for individual SNPs from the association results as reported by the CRP Coronary Heart Disease Genetics Collaboration (CCGC)[55], which represent one unit (in mg/L) increase of the natural log of CRP (lnCRP) per dose of the coded allele. These four tagging SNPs represent 98% of the common variation in the *CRP* gene assuming minor allele frequency ≥0.05 and an r^2 threshold of ≥0.8, and aggregately explain ~2% of the total variation (i.e. phenotypic variance) in serum CRP levels in populations of European descent[55,59]. Second, we constructed a multilocus GRS, named GRS_{GWAS}, that combined 18 SNPs associated with serum CRP levels at genome-wide significance ($P<5\times10^{-8}$; S2 Table and S3 Table), derived from a large meta-GWAS analysis for CRP conducted by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium[30]. This multilocus GRS explains approximately ~5% of the total variation in serum CRP levels[30].

We integrated ω for each CRP SNP from the reference data of CCGC[55] or meta-analysis of GWASs[30] for CRP levels with the summary association statistics extracted from GWAS consortia for each outcome (S1 Data and S2 Methods. This Mendelian randomization approach using meta-GWAS summary statistics data is equivalent to an inverse-variance weighted meta-analysis and has previously been validated in comparison to individual level data[57,60]. To estimate the causal effect of CRP on an outcome, we obtained the β values (estimated effects from regression analysis) for CRP SNPs on the outcome with standard errors of se_β from the corresponding GWAS results. Where no summary statistics for a CRP SNP in the GRS IVs were available in the look-up dataset, we chose a proxy SNP that had the highest linkage disequilibrium (LD) with the initial SNP (r^2>0.9 in HapMap release 22; S3 Table). If several proxy SNPs had the exact same r^2 values, we chose the proxy nearest to the original SNP in the instrument. Separate regressions of outcomes on GRSs were performed to calculate α_{IV} estimators (i.e. causal IV estimator) for each outcome. Correspondingly, the value of a GRS is the sum of the ω values, which is multiplied by the allele dosage (i.e. 0, 1 or 2) for each CRP SNP in the CCGC or in the CHARGE CRP consortium[30,55]. For uncorrelated SNPs, when maximizing the likelihood function, the α_{IV} value and its standard error, se_α, can be approximated with the formula: $\{\alpha \cong (\Sigma\omega \times \beta \times se_\beta^{-2})/(\Sigma\omega^2 \times se_\beta^{-2})\}$ with its $\{se_\alpha \cong \surd 1/ \Sigma\omega^2 \times se_\beta^{-2}\}$. Since lnCRP was used as the outcome in reference studies[3055], to obtain the ω values (i.e. effect sizes) for each of the CRP SNPs, a unit increase in lnCRP equals to a 10 symmetric percentage (s%) increase in CRP levels, which corresponds to a unit change in level of a continuous outcome or logit of risk-estimate (i.e. beta coefficient) for a dichotomous outcome[61]. The α_{IV} value (i.e. causal estimate) for each CRP SNP was, therefore, presented for each outcome as corresponding to a 10 s% increase in actual CRP levels. During the course of this study, an updated larger GWAS dataset for CAD become publicly available (CARDIoGRAMplusC4D Consortium, release 2015[41]), we therefore re-did the analyses for CAD using the 2015 released data.

To assess which SNP might have violated one of the key Mendelian randomization assumptions termed pleiotropy, we performed goodness-of-fit tests to correct both GRSs for heterogeneity of their corresponding SNPs' effects on each outcome. Heterogeneity, which indicates potential presence of pleiotropy, was measured using Q statistics, and was considered statistically significant at a conservative uncorrected *P* value<0.05. Although heterogeneity could be an indicator of pleiotropy; there are other factors that could introduce heterogeneity in the analyses. Therefore, even though the adopted adjustments for heterogeneity that we have taken could be over-conservative, we have taken this method in order to minimize false positives. After stepwise removal of SNPs with potential pleiotropic effects, we repeated the analyses until significant heterogeneity was no longer observed.

To further ensure the strength of these two GRSs as IVs, we generated an *F* statistic for each outcome. We used variance in lnCRP levels explained by each set of CRP SNPs (2% and 5% respectively for GRS_{CRP}, GRS_{GWAS}), to calculate *F* statistics using the formula as F-statistic=$(R^2\times(n-1\text{-}K))/((1-R^2)\times K)$, where "$R^2$" represents proportion of variability in the CRP that is explained by the GRS, "n" represents sample size, and "K" represents number of IVs included in model (i.e. for this study K=1) [62]. As a rule of thumb, an *F* value above 10 indicates that a causal estimate is unlikely to be biased due to weak instruments[57].

Multiple testing

The present study included 32 independent sample-sets. Per each sample-set, we did one statistical test, for which a global nominal significance level of 0.05 would be considered as satisfactory to derive conclusions. The need for correction for multiple testing is debatable. Nevertheless, to ensure the validity of our conclusions, we took a conservative approach, and applied a Bonferroni corrected significance threshold calculated as 0.05 divided by 32 (i.e. 0.0016). We present our results and discussion at three different levels of confidence for corresponding causal estimates; we considered a statistical test with an observed P value more than 0.05 as a definite non-significant result yielded no association; an observed P value equal or less than 0.05 as nominal significant evidence for a potential causal association but yet to be confirmed; and an observed P value equal or less than 0.0016 as statistically significant evidence for a causal association.

CRP polygenic risk score and schizophrenia using individual level data

In an ancillary follow-up study as was inspired by the editors and the reviewers, we aimed further to determine whether CRP_{GWAS} was causally associated with schizophrenia using individual-level data retrieved from the Psychiatric Genomics Consortium (PGC) Schizophrenia dataset (S3 Methods)[54]. This dataset consisted of 36 independent cohorts with a combined 25,629 cases and 30,976 controls for which we had ethics approval (S4 Methods). Three family-based samples of European ancestry (1,235 parent affected-offspring trios) were excluded from our analysis. To evaluate whether the observed protective causal association between CRP_{GWAS} and schizophrenia was persistent, we studied whether the CRP_{PRS} would be also protectively associated with schizophrenia. Briefly, CRP_{PRS} were calculated for each individual by summing the total effect of the SNP dosages by its effect size. In addition to the 18 genome-wide significant CRP SNPs, we grouped sub-threshold CRP-associated SNPs at the following *P*-value thresholds: $1\mathrm{x}10^{-4}$, 0.001, 0.01, 0.05, and 0.1. Standardized CRP_{PRS} were tested for association with schizophrenia case status in each cohort with adjustment for 10 principal components (PCs). A fixed effects inverse variance weighted meta-analysis was performed across all 36 cohorts to obtain the overall effect size estimate as explained in S4 Methods and elsewhere[63]. The variance in schizophrenia case status explained by $\mathrm{CRP}_{\mathrm{PRS}}$ was estimated using the deviation in Nagelkerke's pseudo-R^2 between a null model (which included 10 PCs) to the full model (which included GRSs in addition to 10 PCs), calculated in R using the Functions for medical statistics book with some demographic data (fmsb) R-package (S3 Methods). Similar to previous studies, statistical significance of CRP_{PRS} were estimated based on their logistic regression coefficient[64], and reported CRP_{PRS} ORs correspond to a one SD increase in CRP_{PRS}[65].

RESULTS

Using the GRS_{CRP}, we first tested whether a *CRP* gene determined increase in lnCRP levels was associated with each outcome. In Table 2, the causal effects of lnCRP estimated for each outcome are summarized. We found no heterogeneity in the IV analyses ($\mathrm{P}_{\mathrm{heterogeneity}}$≥0.11 for all outcomes) while the GRS_{CRP} was a strong instrument (*F*≥31). IV analyses provided nominal evidence for potential causal relationships of lnCRP with risk of Crohn's disease (odd ratio [OR] 0.78 [95%CI 0.65-0.94];P<0.009), psoriatic arthritis (1.45 [1.04-2.04];P<0.03), schizophrenia (0.90 [0.82-0.99];P<0.03), and increase in SBP (mean increase 1.23 (0.45-2.01);P<0.002), and DBP (0.70 (0.20-1.19);P<0.006) in mmHg per 10 s% increase in CRP levels. The GRS_{CRP} showed no significant effect on any of the other outcomes (Table 2 and S1 Fig).

Second, the GRS_{GWAS} showed a statistically significant protective effect of lnCRP on the risk of schizophrenia (OR 0.86 [95%CI 0.79-0.94];P<0.0010) per 10 s% increase in CRP levels (Fig 1, Table 3 and S1 Fig). In a follow-up analysis using the individual-level PGC data, we found that genetic risk scores incorporating the same 18 CRP SNPs used to construct

Table 2. The effect of the CRP Genetic Risk Score instrument of four SNPs in CRP (GRS_{CRP}) with somatic and neuropsychiatric outcomes.

Disease / Trait Class	M	N	Effect size (95% CI) *	P-value	P- het	F-value
Autoimmune/Inflammatory						
Celiac Disease	3	15283	0.96 (0.77-1.21)	0.750	0.19	311.86
Inflammatory Bowel Disease (all)	3	47794	0.97 (0.84-1.13)	0.700	0.30	975.35
Crohn's Disease	4	21389	0.78 (0.65-0.94)	0.009	0.25	436.47
Ulcerative Colitis	4	26405	1.10 (0.92-1.31)	0.290	0.92	538.84
Psoriasis Vulgaris	4	8941	1.23 (0.96-1.57)	0.110	0.95	182.43
Psoriatic Arthritis	4	6880	1.45 (1.04-2.04)	0.030	0.92	140.37
Psoriasis Cutaneous	4	4880	1.10 (0.76-1.59)	0.620	0.60	99.55
Rheumatoid Arthritis	4	25702	0.94 (0.77-1.15)	0.550	0.17	524.55
Systemic Lupus Erythematous	3	4651	1.20 (0.80-1.81)	0.380	0.19	94.88
Systemic Sclerosis	3	7518	1.07 (0.78-1.45)	0.680	0.85	153.90
Type 1 Diabetes	2	26890	1.15 (0.90-1.47)	0.260	0.34	548.73
Knee Osteoarthritis	4	24260	0.94 (0.78-1.13)	0.500	0.23	495.06
Cardiovascular						
Coronary Artery Disease	4	184305	1.00 (0.93-1.07)	0.96	0.65	3761.29
Systolic Blood Pressure **	4	69372	1.23 (0.45-2.01)	0.002	0.51	1415.63
Diastolic Blood Pressure **	4	69368	0.70 (0.2-1.19)	0.006	0.68	1415.71
Ischemic Stroke (all types)	4	9520	1.19 (0.93-1.53)	0.160	0.93	194.24
Ischemic Stroke (Cardioembolic)	4	6762	1.02 (0.65-1.58)	0.940	0.96	137.96
Ischemic Stroke (Large Vessel)	4	6816	1.44 (0.93-2.21)	0.100	0.31	139.06
Ischemic Stroke (Small Vessel)	4	6552	1.18 (0.71-1.95)	0.520	0.36	133.06
Metabolic						
Body Mass Index ***	4	123864	-0.017 (-0.06-0.02)	0.410	0.50	2527.82
Type 2 Diabetes	4	22570	1.11 (0.94-1.32)	0.230	0.50	460.57
Chronic Kidney Disease	4	74354	1.04 (0.88-1.22)	0.670	0.90	1517.39
eGFR for creatinine ****	4	74354	0.004 (-0.01-0.02)	0.400	0.88	1517.39
Serum Albumin Levels *****	4	53189	-0.002 (-0.02-0.01)	0.770	0.88	1085.45
Serum Protein Levels *****	4	25537	0.008 (-0.02-0.04)	0.640	0.12	521.12
Neurodegenerative						
Amyotrophic Lateral Sclerosis	2	12263	0.79 (0.60-1.04)	0.090	0.23	258.39
Alzheimer's Disease	2	13020	1.26 (0.89-1.78)	0.200	0.11	265.67
Parkinson's Disease	3	17352	1.00 (0.85-1.17)	0.960	0.33	354.08
Psychiatric						
Autism	3	1566	1.02 (0.97-1.07)	0.380	0.69	31.92
Bipolar Disorder	4	16731	1.17 (0.97-1.42)	0.110	0.49	341.41
Major Depressive Disorder	3	18759	0.98 (0.81-1.18)	0.810	0.86	382.80
Schizophrenia	3	79845	0.90 (0.82- 0.99)	0.030	0.79	1629.45

Abbreviations: M: number of markers used in the genetic instrument; N: number of samples in the disease/ trait meta-analysis; Effect size (95% CI): Effect size (95% CI) per mg/L increase in lnCRP serum levels; P-value: P-value of goodness of fit test; P-het: P-value of heterogeneity of effect test; F-value: F-statistic value for the used genetic instrument.

* For risk of disease, effect size is given in odds ratios, otherwise in the specific units in which the outcome was measured. Derived from the IV causal estimator α.

** Effect size unit is mm Hg per increase in ln serum CRP (mg/L).

*** Effect size unit is 1 standard deviation per ln mg/L increase in serum CRP (the BMI results were inverse normal transformed to a distribution with $\mu = 0$ and $\sigma = 1$).

**** Effect size unit is ml per min per 1.73 m2, per ln mg/L increase in serum CRP.

***** Effect size unit g/dL, per ln mg/L increase in serum CRP.

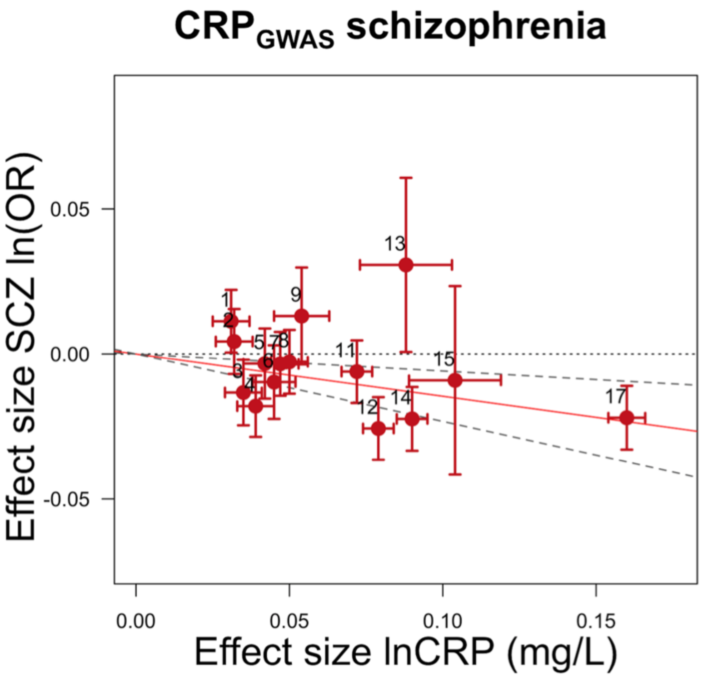


Figure 1. Genetic Risk Score GRS$_{GWAS}$ for schizophrenia. Genetic risk score plots for bipolar disorder and schizophrenia. Horizontal axes: effect size for up to 18 SNPs comprising the GRS$_{GWAS}$ influencing levels of CRP, with corresponding standard error bars. Vertical axes: Log odds ratio for the GRS$_{GWAS}$ SNPs schizophrenia with corresponding standard error bars. The effect estimate of CRP levels on disease risk or trait level is represented by a red solid line with gradient α. The 95% CI of this α estimate is represented by grey dashed lines.

The included SNPs are shown by Arabic numbering as: #1 rs2847281 (gene:*PTPN2*; chr:18;basepair position:12811593); #2: rs340029 (*RORA*;15;58682257); #3 rs6901250(GPRC6A;6;117220718); #4 rs10745954 (*ASCL1*;12;102007224); #5rs4705952(IRF1;5;131867517); #6 rs12037222 (*PABPC4*;1;39837548); #7rs12239046(*NLRP3*;1;245668218); #8 rs6734238 (*IL1F10*;2;113557501); #9rs13233571(*BCL7B*;7;72609167); #10 rs9987289 (*PPP1R3B*;8;9220768); #11 rs1260326 (*GCKR*;2;27584444); #12 rs4129267 (*IL6R*;1;152692888); #13 rs1800961 (*HNF4A*;20;42475778); #14 rs4420065 (*LEPR*;1;6;5934049); #15 rs10521222 (*SALL1*;1;6;49716211); #16 rs1183910 (*HNF1A*;12;119905190); #17 rs2794520 (*CRP*;1;157945440); #18 rs4420638 (*APOC1*;19;50114786).

the GRS$_{GWAS}$ were again significantly associated with a lower risk of schizophrenia (0.96 [0.94-0.98];P<1.72x10^{-6}). This signal persisted when including all SNPs with a less stringent P value threshold of 1x10^{-4} (0.97 [0.95-0.99];P<2.45x10^{-4}). At less stringent P value thresholds, less variance was explained by the logistic model and the protective effect of CRP risk scores became less significant, while across all P value thresholds, the direction of effect was consistently protective (Fig 2 and Fig 3). To ensure that the association between risk alleles for CRP and schizophrenia was not driven by a small number of genome-wide significant SNPs, we performed a leave-one-out sensitivity analysis of the 18 genome-wide SNPs. In the 18 sets of 17 SNPs, the variance explained

Table 3. The effect of the CRP Genetic Risk Score instrument of 18 SNPs associated to CRP (GRS_{GWAS}) with somatic and neuropsychiatric outcomes.

Disease / Trait Class	M	Effect size (95% CI) *	P-value	P-het	F-value
Autoimmune/Inflammatory					
Celiac Disease	18	0.99(0.85-1.16)	0.930	$7.2x10^{-4}$	804.26
Inflammatory Bowel Disease (all)	15	0.85(0.74-0.98)	0.030	$1.4x10^{-5}$	2515.37
Crohn's Disease	17	0.81(0.70-0.94)	0.005	$4.4x10^{-7}$	1125.63
Ulcerative Colitis	17	1.05(0.91-1.21)	0.490	0.01	1389.63
Psoriasis Vulgaris	17	1.12(0.90-1.40)	0.310	0.19	470.47
Psoriatic Arthritis	17	1.36(1.00-1.84)	0.049	0.04	362.00
Psoriasis Cutaneous	17	1.00(0.72-1.39)	0.990	0.16	256.74
Rheumatoid Arthritis	18	0.93(0.80-1.08)	0.350	$1.8x10^{-6}$	1352.79
Systemic Lupus Erythematous	11	1.06(0.71-1.58)	0.780	0.27	244.68
Systemic Sclerosis	11	0.84(0.62-1.14)	0.280	0.63	396.89
Type 1 Diabetes	15	1.10(0.92-1.31)	0.310	$3.47x10^{-3}$	1415.16
Knee Osteoarthritis	18	1.17(1.01-1.36)	0.040	0.10	1276.74
Cardiovascular					
Coronary Artery Disease	18	0.88 (0.84-0.94)	$2.4x10^{-5}$	$7.5x10^{-12}$	9403.21
Systolic Blood Pressure **	18	0.72(0.11-1.34)	0.020	0.14	3650.84
Diastolic Blood Pressure **	18	0.45(0.06-0.84)	0.020	0.02	3651.05
Ischemic Stroke (all types)	18	1.06(0.87-1.29)	0.570	0.37	500.95
Ischemic Stroke (Cardioembolic)	18	0.98(0.69-1.39)	0.920	0.35	355.79
Ischemic Stroke (Large Vessel)	18	1.30(0.92-1.82)	0.140	0.97	358.63
Ischemic Stroke (Small Vessel)	18	0.85(0.58-1.25)	0.420	0.76	343.16
Metabolic					
Body Mass Index ***	18	-0.005(-0.03-0.02)	0.740	0.11	6519.11
Type 2 Diabetes	18	1.090(0.95-1.24)	0.210	$1.8x10^{-3}$	1187.79
Chronic Kidney Disease	18	0.960(0.84-1.09)	0.500	0.07	3913.26
eGFR for creatinine ****	18	0.011(0.003-0.02)	0.005	$7.2x10^{-9}$	3913.26
Serum Albumin Levels *****	18	0.011(0.0004-0.02)	0.041	$2.3x10^{-18}$	2799.32
Serum Protein Levels *****	18	0.031 (0.008-0.05)	0.009	0.03	1343.95
Neurodegenerative					
Amyotrophic Lateral Sclerosis	8	1.01 (0.79-1.29)	0.960	0.56	666.37
Alzheimer's Disease	11	1.26 (0.99-1.61)	0.060	0.23	685.16
Parkinson's Disease	10	1.06 (0.90-1.25)	0.500	0.50	913.16
Psychiatric					
Autism	9	0.89 (0.70-1.13)	0.350	0.99	82.32
Bipolar Disorder	18	1.21 (1.05-1.40)	0.007	0.15	880.47
Major Depressive Disorder	15	1.14 (0.96-1.36)	0.140	0.84	987.21
Schizophrenia	15	0.86 (0.79-0.94)	0.001	0.66	4202.26

Abbreviations: M: number of markers used in the genetic instrument; *Effect size (95% CI)*: Effect size (95% CI) per mg/L increase in lnCRP serum levels; *P-value*: P-value of goodness of fit test; *P-het*: P-value of heterogeneity of effect test; *F-value*: F-statistic value for the used genetic instrument

* For risk of disease, effect size is given in odds ratios, otherwise in the specific units in which the outcome was measured. Derived from the IV causal estimator α.

** Effect size unit is mm Hg per increase in ln serum CRP (mg/L).

*** Effect size unit is 1 standard deviation per ln mg/L increase in serum CRP (the BMI results were inverse normal transformed to a distribution with $\mu = 0$ and $\sigma = 1$).

**** Effect size unit is ml per min per 1.73 m2, per ln mg/L increase in serum CRP.

***** Effect size unit g/dL, per ln mg/L increase in serum CRP.

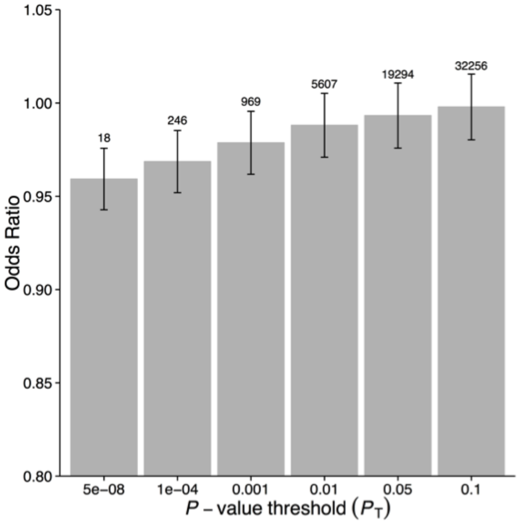


Figure 2. Polygenic risk scores for elevated CRP levels and protective effect on schizophrenia, using individual level genetic data.

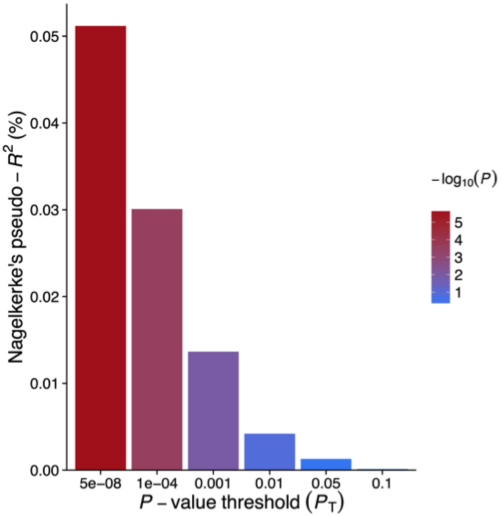


Figure 3. **Polygenic risk scores for elevated CRP levels and explained variance of schizophrenia using individual level genetic data.**

(Nagelkerke's pseudo-R^2) ranged from 0.012% to 0.034%, with P values ranging from 9.3x10^{-5} to 1.6x10^{-2}, suggesting that the protective effect observed between risk alleles for CRP and schizophrenia was not driven by a small number of SNPs with large effects.

The GRS$_{GWAS}$ also showed moderate but nominally significant effects of lnCRP on the risk of IBD (OR 0.85 [95%CI 0.74-0.98];P<0.03), Crohn's disease (0.81[0.70-0.94];P<0.005), psoriatic arthritis (1.36[1.00-1.84];P<0.049), knee osteoarthritis (1.17[1.01-

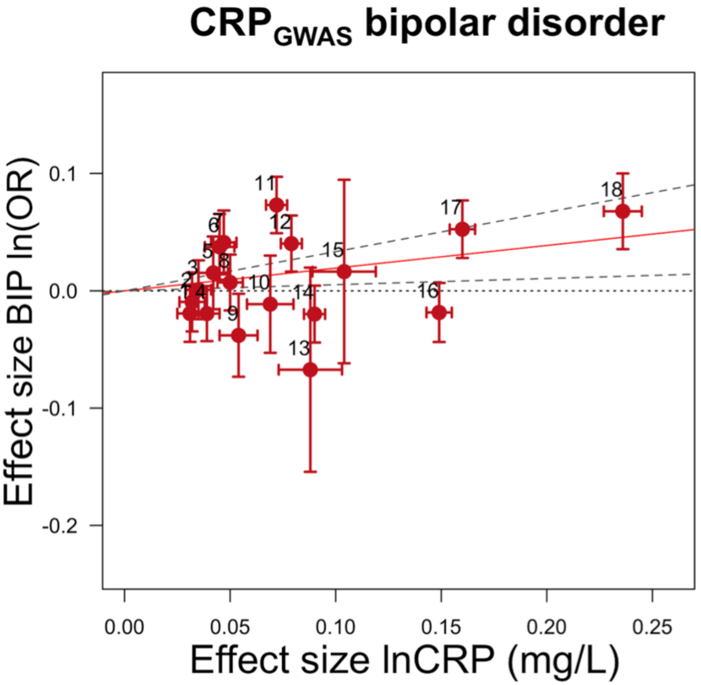


Figure 4. **Genetic Risk Score GRS_{GWAS} for bipolar disorder.** Genetic risk score plots for bipolar disorder and schizophrenia. Horizontal axes: effect size for up to 18 SNPs comprising the GRS_{GWAS} influencing levels of CRP, with corresponding standard error bars. Vertical axes: Log odds ratio for the GRS_{GWAS} SNPs schizophrenia with corresponding standard error bars. The effect estimate of CRP levels on disease risk or trait level is represented by a red solid line with gradient α. The 95% CI of this α estimate is represented by grey dashed lines.

The included SNPs are shown by Arabic numbering as: #1 rs2847281 (gene:*PTPN2*; chr:18;basepair position:12811593); #2: rs340029 (*RORA*;15;58682257); #3 rs6901250(GPRC6A;6;117220718); #4 rs10745954 (*ASCL1*;12;102007224); #5rs4705952(IRF1;5;131867517); #6 rs12037222 (*PABPC4*;1;39837548); #7rs12239046(*NLRP3*;1;245668218); #8 rs6734238 (*IL1F10*;2;113557501); #9rs13233571(*BCL7B*;7;72609167); #10 rs9987289 (*PPP1R3B*;8;9220768); #11 rs1260326 (*GCKR*;2;27584444); #12 rs4129267 (*IL6R*;1;152692888); #13 rs1800961 (*HNF4A*;20;42475778); #14 rs4420065 (*LEPR*;1;6;5934049); #15 rs10521222 (*SALL1*;1;6;49716211); #16 rs1183910 (*HNF1A*;12;119905190); #17 rs2794520 (*CRP*;1;157945440); #18 rs4420638 (*APOC1*;19;50114786).

1.36];P<0.04), and bipolar disorder (1.21[1.05-1.40];P<0.007) while it was statistically significant for coronary artery disease (CAD) (0.88 [0.84-0.94];P<$2.4x10^{-5}$), (Table 3, Fig 4 and S1 Fig). GRS_{GWAS} revealed a nominally significant increase of 0.72 (95%CI 0.11-1.34;P<0.02), and 0.45 (0.06-0.84;P<0.02) mmHg in SBP and DBP respectively (Table 3, S1 Fig). Likewise, genetically 10 s% increase in CRP levels was nominally associated with a 0.01 ml/min/1.73m^2 (0.003-0.02;P<0.005) higher estimated glomerular filtration rate from serum creatinine ($eGFR_{cr}$), 0.01 g/dl (0.0004-0.02;P<0.04) higher albumin, and 0.03 g/dl (0.008-0.05;P<0.009) higher serum protein levels. The remaining outcomes tested for causal associations using GRS_{GWAS} did not reach statistical significance, though the corresponding GRS_{GWAS} proved to be a strong IV with *F* values ≥82 (Table 3; S1 Fig).

Using the GRS_{GWAS}, there was no significant evidence of heterogeneity of the effect sizes for knee osteoarthritis, bipolar disorder, schizophrenia, and SBP, while the heterogeneity test was statistically significant for psoriatic arthritis, IBD, Crohn's disease, CAD, DBP, eGFR_{cr}, serum albumin and serum protein. These heterogeneities in the effects of GRS_{GWAS} may be attributable to pleiotropic effects of SNPs used to build the GRS_{GWAS}. We subsequently performed a stepwise removal of SNPs from GRS_{GWAS} until no significant heterogeneity remained and presented the results in Table 4. This adjustment in the GRS_{GWAS} resulted in the removal of three SNPs from the GRS_{GWAS} for IBD (in *GCKR*, *IRF1*, *PTPN2*), five SNPs for Crohn's disease (in *GCKR*, *IL6R*, *IRF1*, *PABPC4*, *PTPN2*), one SNP for psoriatic arthritis (in *IRF1*), three SNPs for CAD (in *APOC1, HNF1A, IL6R*), one SNP for DBP (in *PABPC4*), two SNPs for eGFR (in *LEPR* and *GCKR*), six SNPs for serum albumin (in *APOC1*, *BCL7B*, *GCKR, PPP1R3B*, *PTPN2*, *IRF1*), and one SNP for serum protein levels (in *GCKR*). After removal of these variants from the GRS_{GWAS}, we found no statistically significant (at $P<0.0016$) association between genetically increased lnCRP levels and any of these outcomes (Table 4). However, the effect estimate of CRP on DBP, serum albumin, and psoriasis arthritis showed nominal association at $P<0.05$). For example in DBP, 17 SNPs remained in the GRS_{GWAS} and yielded a slightly lower causal estimate (when compared to the values before adjustment) of 0.39 (-0.01 to 0.78) mmHg increase in DBP per 10 s% increase in lnCRP levels with a nominal significance of $P<0.05$.

Likewise, we hypothesized that a non-significant effect of CRP using GRS_{GWAS} on celiac disease, ulcerative colitis, rheumatoid arthritis, type 1 diabetes and type 2 diabetes can be to some extent explained by significant heterogeneity observed for these outcomes (Table 3). This adjustment in the GRS_{GWAS} resulted in the removal of two SNPs from the GRS_{GWAS} for celiac disease (in *PABPC4*, *PTPN2*), one SNP for ulcerative colitis (in *GCKR*), five SNPS for rheumatoid arthritis (in *HNF4A, IL6R*, S*ALL1, NLRP3, PTPN2),* one SNP for type 1 diabetes (in *PTPN2*), and one SNP for type 2 diabetes (in *APOC1*). After adjusting for heterogeneity, the association of GRS_{GWAS} with these outcomes remained statistically non-significant (Table 4).

Table 4. The effect of the CRP Genetic Risk Score instrument of 18 SNPs associated to CRP (GRS$_{GWAS}$) with somatic and neuropsychiatric outcomes after correcting for heterogeneity.

Disease / Trait Class	M	Effect size (95% CI)*	P-value	P- het
Autoimmune/Inflammatory				
Celiac Disease	16	1.05 (0.90-1.23)	0.56	0.10
Inflammatory Bowel Disease	12	0.92 (0.79-1.06)	0.24	0.14
Crohn's Disease	12	0.93 (0.79-1.08)	0.34	0.12
Ulcerative Colitis	16	1.11 (0.96-1.28)	0.16	0.12
Psoriatic Arthritis	16	1.42 (1.05-1.94)	0.02	0.14
Rheumatoid Arthritis	13	0.83 (0.71-0.97)	0.02	0.09
Type 1 Diabetes	14	1.06 (0.89-1.27)	0.52	0.07
Cardiovascular				
Coronary Artery Disease	15	0.98 (0.91-1.06)	0.65	0.20
Diastolic Blood Pressure **	17	0.385 (-0.008-0.78)	0.05	0.09
Metabolic				
Type 2 Diabetes	17	0.95 (0.82-1.10)	0.52	0.09
eGFR for creatinine ***	16	0.001 (-0.007-0.01)	0.74	0.11
Serum Albumin ****	12	-0.017 (-0.029- -0.004)	0.01	0.07
Serum Protein ****	17	0.021 (-0.002-0.05)	0.07	0.31

Abbreviations: M: number of markers used in the genetic instrument; Effect size (95% CI): Effect size (95% CI) per mg/L increase in lnCRP serum levels; P-value: P-value of goodness of fit test; P-het: P-value of heterogeneity of effect test.

* For risk of disease, effect size is given in odds ratios, otherwise in the specific units in which the outcome was measured. Derived from the IV causal estimator α.

** Effect size unit is mm Hg per increase in ln serum CRP (mg/L).

*** Effect size unit is ml per min per 1.73 m2, per ln mg/L increase in serum CRP.

**** Effect size unit g/dL, per ln mg/L increase in serum CRP

DISCUSSION

In this large scale cross-consortia Mendelian randomization study of 32 complex outcomes, we found evidence for a potential protective causal relationship between elevated CRP levels, and schizophrenia in both genetic IVs (i.e. GRS$_{CRP}$ and GRS$_{GWAS}$), and confirmed this protective relationship in follow-up analyses using individual-level genotype data from the schizophrenia GWAS. We also found statistically significant association with CAD, and nominally significant evidence for a predisposing causal association of CRP levels with IBD, Crohn’s disease, psoriasis arthritis, knee osteoarthritis, SBP, DBP, eGFR, albumin and serum protein levels, and bipolar disorder using GRS$_{GWAS}$ as an IV. However, after adjustment for heterogeneity, neither GRS showed a significant effect (at $p<0.0016$) on any of these outcomes, including CAD, nor on the other 20 other common somatic and psychiatric outcomes, including celiac disease, ulcerative colitis, psoriasis (all types), rheumatoid arthritis, systemic lupus erythematous, systemic sclerosis, type 1 and 2 diabetes, stroke (all types), BMI, chronic kidney disease, amyotrophic lateral sclerosis, Alzheimer's disease, Parkinson's disease, autism, and major depressive disorder.

CRP protection against schizophrenia

Strikingly, as opposed to current literature and previously inconclusive small scale studies[66,67,68], our finding suggest that genetically elevated levels of CRP are not predisposing but in fact protective for schizophrenia. The significant causal protective role of CRP with schizophrenia was consistent in both IVs using summary statistics i.e. GRS_{CRP} and GRS_{GWAS}. When incorporating 18 genome-wide CRP-associated SNPs using individuals level data, we confirmed a modest, but significant, protective effect for schizophrenia. This signal persisted when including all SNPs with a less stringent *P*-value threshold of 1×10^{-4}. Notably, the leave-one-out sensitivity analysis revealed that the genetic overlap between CRP levels and schizophrenia we observed at genome-wide and 1×10^{-4} significance thresholds was not driven by few major SNPs. In contrast, others have previously shown that CRP levels are significantly elevated in patients with schizophrenia[69,70]; with a recent meta-analysis concluding that the association between elevated CRP and schizophrenia is indeed robust[71]. Given that clinical studies report elevated CRP levels in schizophrenia one would expect to find that raising alleles for elevated CRP would confer an increased risk for schizophrenia. The fact that we found a completely opposite effect—indeed in a cohort of over 25,000 cases and 30,000 controls—should give pause when deriving clinical meaning from these results. Our observation that a genetically determined marginal increase in the levels of CRP is likely to be protective for schizophrenia, may fuel the debate about whether the observed CRP elevation in schizophrenia is a by-product of the pathogenesis of schizophrenia or directly contributing to clinical manifestations of the disorder[6]. This finding may also point out potential biases in previous studies regarding the causes of elevated CRP levels in patients with schizophrenia such as reverse causality and/or pleiotropic effects within chosen instruments.

The exact mechanism on how elevated CRP levels are linked to schizophrenia requires a well-defined experimental analysis. In addition to CRP variants other recent studies have identified several inflammatory genetic variants associated to schizophrenia, and bipolar disorder, which includes variants in major histocompatibility complex (MHC) region on chromosome 6p21 [72], harboring many cytokine genes[54,73–76], in the *IL10* promoter[77], *TNF* promoter[78], *IL-1B*[79] and *C4*[80].

Biological annotation

Following the comments made by the reviewers, we explored the possible underlying pathways which may explain the potential protective causal association between CRP and schizophrenia. We performed a follow *insilico* functional pathway analysis using previously reported approach[81] as summarized in S5 Methods and in Table S4 through S13 Table. In brief, our results show that pathways associated with the interferon response are significantly enriched amongst genes harboured by CRP loci and their associated eQTLs, as well as differentially expressed genes between schizophrenia cases and controls. Previous studies showed that the induction of T-cell IFN cytokine release stimulates microglia and astrocytes to facilitate glutamate clearance in neuronal cells without evoking inflammatory mediators[82,83]. One could speculate that CRP-interferon pathways may induce neuroprotection by contributing to glutamate clearance, leading to the protection of neurons against oxidative stress associated with excess of glutamate[84,85], and therefore offering a protective effect against schizophrenia.

CRP GRS$_{GWAS}$ association to bipolar disorder

As for bipolar disorder, we found a nominal effect of 1.21 increase in risk for bipolar disorder by a 10 s% increase in CRP levels. Though this nominal predisposing effect needs to be confirmed, our finding corroborates epidemiological observations suggesting that elevated CRP is associated with the disease and support a potential causal influence of general inflammation in bipolar disorder[86]. We note that, though it may be biologically sensible, this result failed to pass multiple testing correction. Confirmation by replication in independent cohorts, functional follow-up analyses, and/or the use of a stronger CRP GRS$_{GWAS}$ from upcoming studies are required to make a factitive conclusion.

CRP GRS$_{GWAS}$ association to blood pressure and hypertension

We found nominally significant evidence for up to ~0.70 mmHg increased blood pressure for a 10 s% increase in CRP levels and no evidence for heterogeneity for SBP. Additionally, there was nominally borderline significance of a causal association between CRP and DBP after adjustment for heterogeneity. These nominally significant findings, on the one hand, are in line with numerous epidemiological studies that have highlighted an association between elevated CRP and an increased risk of hypertension. For instance, one study found an association between CRP loci and hypertension in Asians[87]. An additional line of support for a possible causal association of CRP and blood pressure comes from an experimental study where an increase of *CRP* gene expression in mice, and subsequently, CRP protein levels, led to a rise in particularly SBP[88]. Moreover an ex-vivo study by Zhou *et al.* has shown that combining IL-6 treatment and mechanical strain has led to a consistent increase in CRP expression at protein and mRNA levels in smooth muscle cells[89]. Both inflammatory factors and local mechanical strains are abundant in blood vessels and are well known risk factors for high blood pressure. However, on the other hand, our finding did not reach a statically significant level after correction for multiple testing; thus it may echo previous Mendelian randomization studies which have failed to find a causal relationship between CRP levels and blood pressure or hypertension in Europeans[90,91]. However our systematic literature review showed previous studies had some limitations (S1 Table). For instance, no study used a refined GWAS set of 18 CRP associated SNPs instead they tested single or a limited set of CRP SNPs. Using such instruments might have led to biased estimates as their corresponding effects on CRP levels have been found to be small[30,57]. A combination of weak instruments and low sample sizes might lead to type II error[28,57], and hence to a conclusion of no causal association between CRP and blood pressure traits in previous studies. Taken together, a direct link between CRP and blood pressure remains to be elucidated, though our nominal association between $\text{GRS}_{\mathit{CRP,}}$ $\text{GRS}_{\mathit{GWAS}}$ and blood pressure do add to a line of findings from experimental studies suggesting a potential causal relationship between CRP and blood pressure.

CRP $\text{GRS}_{\mathit{GWAS}}$ association to osteoarthritis

Our nominally significant findings that CRP might be a potential causal factor for knee osteoarthritis (using $\text{GRS}_{\mathit{GWAS}}$), should be interpreted with caution. In line with our findings, we have previously shown earlier that levels of CRP were higher in women with early radiological knee osteoarthritis (i.e. Kellgren-Lawrence grade 2+), and in women whose disease progressed[92]. Additionally, another study showed that genetically elevated CRP levels contribute to osteoarthritis severity[93]. However, other studies have found contrasting results[71,72,94]. One systematic review provided evidence that the relationship between CRP and osteoarthritis does exist, but is dependent on BMI[95]. It remains to be further investigated whether weight gain over the lifetime mediates the potential causal association between genetically elevated CRP and knee osteoarthritis.

CRP GRS$_{GWAS}$: no association to other remaining outcomes

The present study was able to calculate nominal causal estimates for IBD, Crohn's disease, psoriatic arthritis, CAD, eGFRcr, serum albumin or protein using a CRP GRS$_{GWAS}$; but they were altered by removal of SNPs from GRS$_{GWAS}$ based on heterogeneity tests resulting in nominal, or non-significant associations. These outcomes appeared therefore to have heterogeneity in causal estimates suggesting these observed estimates were biased likely due to pleiotropic effects of CRP loci. These results corroborate negative findings of previous studies (S1 Table), suggesting a causal role of CRP in these traits and diseases is unlikely.

Methodological concerns and advantages

Pleiotropic biases in MR analyses using CRP GRS$_{GWAS}$

A detailed evaluation of pleiotropic SNPs in our study showed that the applied method to identify heterogeneity sources was able to indicate and exclude several already known pleiotropic loci from the GRS$_{GWAS}$ IV. For instance, the use of a SNP in *IL6R* (rs4129267) amongst others resulted in heterogeneity of effects on CAD risk. The same variant contributed to heterogeneity of effects for Crohn's disease in our study, and it has been shown that this SNP is associated with levels of biomarkers other than CRP[56]. Further, a Mendelian randomization study found that *IL6R* SNPs, specifically, the non-synonymous SNP rs8192284, are associated with CAD risk and CRP levels[96]. Our selected *IL6R* SNPs, namely rs4537545 or rs4129267, are in extremely high LD with rs8192284 ($r^2 \geq 0.96$ for both SNPs in Hapmap data, CEU population). Carriers of the risk allele of rs8192284 have higher CRP, IL6 and fibrinogen levels[96]. Fibrinogen is also a well-known risk factor for CAD. Therefore, it is unclear so far which biomarker(s) mediates the effect of *IL6R* SNPs on CAD. Besides the *IL6* locus, *APOC1 and PABPC4* have been indicated as pleiotropic in three, and *PTPN2* and *GCKR* in six out of 32 our investigated outcomes. Taken together, we were able to disentangle at least part of the pleiotropy regarding the causal estimates of CRP for outcomes. Again, we found no significant association of CRP GRS$_{GWAS}$ after adjustment for heterogeneity.

Using summary statistics of large scale consortia

It is of utmost interest whether the observed effect of CRP as a risk predictor for human disease is causal, and thus whether reduction of CRP levels will lower the risk of disease. Here, we investigated the causality of CRP in 32 phenotypes by leveraging very large samples sizes collected by GWAS consortia, an approach that was much better powered than most previous Mendelian randomization studies. We found that genetically elevated CRP levels approximated by powerful instruments did not appear to contribute directly to most of the studied somatic and psychiatric outcomes. Our findings are consistent with previous Mendelian randomization studies reporting null associations of genetically elevated CRP levels with inflammation-related outcomes including CAD[56,59,97], type 2 diabetes[98], BMI[99], Alzheimer's disease and depression[100]. All previous Mendelian randomization studies were substantially limited to a single or a few outcomes, used only SNPs in the *CRP* gene or had sample sizes much smaller than the present study (S1 Table). In addition to these studies, the use of current GWAS data do not corroborate epidemiological observations suggesting that elevated CRP levels are associated with amyotrophic lateral sclerosis[101], Alzheimer's disease[102], Parkinson's disease[103], and major depressive disorder[104]. Furthermore, patients with immunity-related disorders frequently have a very high CRP level (as high as 100 mg/L) due to their disease status. Our findings may therefore more favorably indicate reverse causality. Taken together, we showed that CRP is highly unlikely to contribute causally to most of the major common somatic and neuropsychiatric outcomes that are investigated in the present study, with the possible exception of schizophrenia.

Strength of Instrumental Variables

Results presented in Table 2 show that our GRS_{CRP} is not a weak instrument, as indicated by its high F values owing to the large sample sizes of available outcomes GWASs for the phenotypes under study. The strength of our instruments increased considerably in all disease classes when we used variants of multiple loci associated with CRP in GWAS. However, the variants comprising the CRP GRS_{GWAS} explain on average only a moderate ~5% of the total variance in baseline CRP levels[30]. Moreover, the possibility of effect modification by non-genetic CRP related factors on the outcomes remains to be investigated. We may be able to create even stronger instruments based on ongoing efforts to identify additional variation influencing CRP levels. Even if larger sample sizes and stronger instruments can be realized, the overwhelming lack of causal effects observed for most outcomes in our study implies that therapies targeted at lowering CRP will not directly result in decreased risk of the investigated outcomes, or a better symptom management [105,106].

Using summary statistics instead of individual level data

Here we used summary associations statistics obtained from previously conducted meta-GWASs in order to maximize our study power. One may argue this may induce bias, compared to when one uses the individual level data. Nevertheless, previous studies showed high agreement in results from GWAS summary data and individual level data Mendelian randomization methods[60,107]; Furthermore, our analyses of individual-level data for schizophrenia led to the same conclusion as our analyses using summary statistics data, confirming the robustness of our methodological approach.

Other potential sources of biases

An important rationale for Mendelian randomization is that the gene variants do not change over time and are inherited randomly. Thus, the genetic variants are considered free from confounding and reverse causation[108]. However, one cannot completely control for the possibility of confounding of genotype—intermediate phenotype—disease associations. For instance, there might be a chance as recognized confounding by ethnic/racial group (i.e. population stratification) is unlikely to be a major problem in most situations[108]. In the present study, we included summary statistics data from highly credible results of meta-GWASs. All the original meta-GWASs have corrected for population stratification at cohort level analyses, and at meta-GWAS level.

Another caveat of MR is that developmental compensation might occur through a genotype being expressed during fetal development which in turn buffer the effects of either environmental or genetic factors, called canalization[108,109]. Therefore, buffering mechanisms could hamper the associations between genetic variants and the outcome of interest. As opposed to this, a lifetime exposure to a risk factor may enhance its effects on the disease[109]. However, it is not clear to what extent genetically determined small changes in any given exposure would be sufficient to induce compensation[108]. All the 32 meta-GWASs from which instrument summary estimates were taken, have been performed in Caucasians, particularly in European and US populations including thousands of samples for each outcome (S1 Table), and so was our previous CRP meta-GWAS from which we have chosen the CRP-associated SNPs to calculate $\text{CRP}_{\textit{GWAS}}$ GRS. Therefore, the results of this MR study are applicable to Caucasians, and are not necessarily generalizable to other ethnic groups.

CONCLUSION

We showed that elevated CRP levels driven by genetic factors are causally associated with protection against schizophrenia, suggesting that CRP may be one important puzzle piece that leads to an improved understanding of the pathogenesis of schizophrenia. We observed nominal evidence that genetically elevated CRP is causally associated with SBP, DBP, knee osteoarthritis and bipolar disorder. Based on current GWAS data, we cannot verify any causal effect of CRP on the other 27 common somatic and neuropsychiatric outcomes investigated in the present study. Therefore, disease associated rise in CRP levels may be interpreted as a response to the disease process rather than a cause for these 27 outcomes. This implies that interventions lowering CRP levels are unlikely to result in decreased risk for the majority of common complex outcomes.

Acknowledgements

We thank for Dan Arking for providing the genome-wide summary statistics data on CRP polymorphisms for the Autism GWAS. We also would like to acknowledge the Rheumatoid Arthritis and Systemic Lupus Erythematosus consortia, the Genetic Investigation of ANthropometric Traits consortium (GIANT), and the Psychiatric Genomics Consortium (PGC) who made their data publicly available from which we extracted summary association statistics for CRP SNPs. The PAGE consortium is grateful to Alex Tsoi for adding the consortium information and membership list. The Type 1 Diabetes consortium would like to thank all participating subjects and families, and also thank Smari Kristinsson, Larus Arni Hermannsson, and Asbjörn Krisbjörnsson of Raförninn for their extensive software design and contribution. We would like to thank people with MND and their families for their participation in this project. We acknowledge sample management undertaken by Biobanking Solutions funded by the Medical Research Council at the Centre for Integrated Genomic Medical Research, University of Manchester.

Consortia co-authors

Please refer to S1 Consortia full name and affiliations coauthors of five consortia including Schizophrenia Working Group of the Psychiatric Genomics Consortium, Autism Spectrum Disorder Working group of the Psychiatric Genomics Consortium, GERADE1, ICBP and PAGE.

Supplementary material

Supplementary Material is available at PLOS Medicine online.

It contains the following :

- S1 Table. Previous Mendelian Randomization Analyses using CRP variants as instruments
- S2 Table. CRP lead variants used in the genetic risk score as instrumental variables
- S3 Table. Proxy SNPs of CRP lead variants used in the genetic risk scores as instrumental variables
- S4 Table. Biologically prioritized candidate gene set associated with CRP used as the input query to the gene set enrichment analysis
- S5 Table. Pathway enrichment results for biologically prioritized candidate gene set associated with CRP used as the input query to the gene set enrichment analysis
- S6 Table. 144 genes that were significantly differentially expressed between schizophrenia and unaffected controls in the hippocampus
- S7 Table. Pathway enrichment results for 144 genes that were significantly differentially expressed between schizophrenia and unaffected controls in the hippocampus
- S8 Table. Biologically prioritized candidate gene set of CRP from Vaez et.al, 2015 (from S5, CRP genes, in blue) and differentially expressed genes in schizophrenia cases vs controls from Hwang et. al 2013 (from S7, SCZ expressed genes, in red) used as the input query to the pathway analysis.
- S9 Table. Pathway enrichment results for the combined set of biologically prioritized candidate gene set of CRP from Vaez et.al, 2015 and differentially expressed genes in schizophrenia cases vs controls from Hwang et. al 2013
- S10 Table. List of genes in 108 genome-wide significant loci associated with schizophrenia
- S11 Table. Brain and blood eQTL for credible sets of SNPs of the 108 schizophrenia loci

- S12 Table. List of genes in 108 genome-wide significant loci associated with schizophrenia (yellow), brain eQTL(red), and blood eQTL(blue) from S.11 and S.12
- S13 Table. Pathway enrichment results for the list of genes in 108 genome-wide significant loci associated with schizophrenia (yellow), and those with brain- (red), and blood eQTL (blue) from S.13
- S1 Consortia. Consortia co-authors and collaborators
- S1 Data. Individual association summary statistics of CRP lead SNPs and/or proxies with traits and diseases
- S1 Methods. Linkage disequilibrium of the four GRSCRP SNPs
- S2 Methods. CRP GRSGWAS in AD and BMI.
- S3 Methods. Web Links
- S4 Methods. CRP Polygenic risk score (CRPPRS) in Schizophrenia.
- S5 Methods. In-silico (gene) pathway analyses highlight the role of IFN in the causal pathway between CRP and SCZ
- S1 Fig. GRSCRP and GRSGWAS per each studied outcome.
- S1 Financial Disclosure. Authors' funding information

REFERENCES

1. Stockinger, B. Immunology: Cause of death matters. Nature 458, 44–45 (2009).
2. Baumgart, D. C. & Carding, S. R. Inflammatory bowel disease: cause and immunobiology. Lancet 369, 1627–1640 (2007).
3. Bruunsgaard, H., Pedersen, M. & Pedersen, B. K. Aging and proinflammatory cytokines. Curr. Opin. Hematol. 8, 131–136 (2001).
4. Cesari, M. et al. Inflammatory markers and cardiovascular disease (The Health, Aging and Body Composition [Health ABC] Study). Am. J. Cardiol. 92, 522–528 (2003).
5. Couzin-Frankel, J. Inflammation bares a dark side. Science 330, 1621 (2010).
6. Fan, X., Goff, D. C. & Henderson, D. C. Inflammation and schizophrenia. Expert Rev Neurother 7, 789–796 (2007).
7. Grimaldi, M. P. et al. Genetics of inflammation in age-related atherosclerosis: its relevance to pharmacogenomics. Ann. N. Y. Acad. Sci. 1100, 123–131 (2007).
8. Hanson, D. R. & Gottesman, I. I. Theories of schizophrenia: a genetic-inflammatory-vascular synthesis. BMC Med. Genet. 6, 7 (2005).
9. Johnson, T. E. Recent results: biomarkers of aging. Exp. Gerontol. 41, 1243–1246 (2006).
10. Kiecolt-Glaser, J. K. & Glaser, R. Depression and immune function: central pathways to morbidity and mortality. J Psychosom Res 53, 873–876 (2002).
11. Kiecolt-Glaser, J. K., McGuire, L., Robles, T. F. & Glaser, R. Emotions, morbidity, and mortality: new perspectives from psychoneuroimmunology. Annu Rev Psychol 53, 83–107 (2002).
12. Lynch, M. A. & Mills, K. H. G. Immunology meets neuroscience--opportunities for immune intervention in neurodegenerative diseases. Brain Behav. Immun. 26, 1–10 (2012).
13. Meyer, U., Schwarz, M. J. & Müller, N. Inflammatory processes in schizophrenia: a promising neuroimmunological target for the treatment of negative/cognitive symptoms and beyond. Pharmacol. Ther. 132, 96–110 (2011).
14. Pardo, C. A., Vargas, D. L. & Zimmerman, A. W. Immunity, neuroglia and neuroinflammation in autism. Int Rev Psychiatry 17, 485–495 (2005).
15. Sansoni, P. et al. The immune system in extreme longevity. Exp. Gerontol. 43, 61–65 (2008).
16. Hotamisligil, G. S. Inflammation and metabolic disorders. Nature 444, 860–867 (2006).
17. Mantovani, A., Allavena, P., Sica, A. & Balkwill, F. Cancer-related inflammation. Nature 454, 436–444 (2008).
18. Dantzer, R., O'Connor, J. C., Freund, G. G., Johnson, R. W. & Kelley, K. W. From inflammation to sickness and depression: when the immune system subjugates the brain. Nat. Rev. Neurosci. 9, 46–56 (2008).
19. Emerging Risk Factors Collaboration et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. Lancet 375, 132–140 (2010).
20. Emerging Risk Factors Collaboration et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. N. Engl. J. Med. 367, 1310–1320 (2012).
21. Wang, X. et al. Inflammatory markers and risk of type 2 diabetes: a systematic review and meta-analysis. Diabetes Care 36, 166–175 (2013).

22. Henriksen, M. et al. C-reactive protein: a predictive factor and marker of inflammation in inflammatory bowel disease. Results from a prospective population-based study. Gut 57, 1518–1523 (2008).

23. Rhodes, B. et al. A genetic association study of serum acute-phase C-reactive protein levels in rheumatoid arthritis: implications for clinical interpretation. PLoS Med. 7, e1000341 (2010).

24. Ridker, P. M. High-sensitivity C-reactive protein as a predictor of all-cause mortality: implications for research and patient care. Clin. Chem. 54, 234–237 (2008).

25. Dehghan, A. et al. Genetic variation, C-reactive protein levels, and incidence of diabetes. Diabetes 56, 872–878 (2007).

26. Lee, C. C. et al. Association of C-reactive protein with type 2 diabetes: prospective analysis and meta-analysis. Diabetologia 52, 1040–1047 (2009).

27. Danesh, J. & Pepys, M. B. C-reactive protein and coronary disease: is there a causal link? Circulation 120, 2036–2039 (2009).

28. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med 27, 1133–1163 (2008).

29. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC) et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ 342, d548 (2011).

30. Dehghan, A. et al. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123, 731–738 (2011).

31. Dubois, P. C. A. et al. Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 42, 295–302 (2010).

32. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. 42, 1118–1125 (2010).

33. Anderson, C. A. et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat. Genet. 43, 246–252 (2011).

34. Nair, R. P. et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat. Genet. 41, 199–204 (2009).

35. Ellinghaus, E. et al. Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. Nat. Genet. 42, 991–995 (2010).

36. Stahl, E. A. et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42, 508–514 (2010).

37. Hom, G. et al. Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N. Engl. J. Med. 358, 900–909 (2008).

38. Radstake, T. R. D. J. et al. Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. Nat. Genet. 42, 426–429 (2010).

39. Bradfield, J. P. et al. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. PLoS Genet. 7, e1002293 (2011).

40. Ramos, Y. F. M. et al. Meta-analysis identifies loci affecting levels of the potential osteoarthritis biomarkers sCOMP and uCTX-

II with genome wide significance. J. Med. Genet. 51, 596–604 (2014).

41. Nikpay, M. et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. 47, 1121–1130 (2015).

42. International Consortium for Blood Pressure Genome-Wide Association Studies et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478, 103–109 (2011).

43. International Stroke Genetics Consortium (ISGC) et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. Nat. Genet. 44, 328–333 (2012).

44. Speliotes, E. K. et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. 42, 937–948 (2010).

45. Voight, B. F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. 42, 579–589 (2010).

46. Köttgen, A. et al. New loci associated with kidney function and chronic kidney disease. Nat. Genet. 42, 376–384 (2010).

47. Franceschini, N. et al. Discovery and fine mapping of serum protein loci through transethnic meta-analysis. Am. J. Hum. Genet. 91, 744–753 (2012).

48. Shatunov, A. et al. Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. Lancet Neurol 9, 986–994 (2010).

49. Hollingworth, P. et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat. Genet. 43, 429–435 (2011).

50. International Parkinson Disease Genomics Consortium et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet 377, 641–649 (2011).

51. Weiss, L. A., Arking, D. E., Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly, M. J. & Chakravarti, A. A genome-wide linkage and association scan reveals novel loci for autism. Nature 461, 802–808 (2009).

52. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. 43, 977–983 (2011).

53. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium et al. A mega-analysis of genome-wide association studies for major depressive disorder. Mol. Psychiatry 18, 497–511 (2013).

54. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427 (2014).

55. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC) et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ 342, d548 (2011).

56. Elliott, P. et al. Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. JAMA 302, 37–48 (2009).

57. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet. Epidemiol. 37, 658–665 (2013).

58. Dastani, Z. et al. Novel loci for adiponectin levels and their influence on type 2 diabetes

and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. PLoS Genet. 8, e1002607 (2012).

59. Zacho, J. et al. Genetically elevated C-reactive protein and ischemic vascular disease. N. Engl. J. Med. 359, 1897–1908 (2008).

60. Voight, B. F. et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380, 572–580 (2012).

61. Cole, T. J. Sympercents: symmetric percentage differences on the 100 log(e) scale simplify the presentation of log transformed data. Stat Med 19, 3109–3125 (2000).

62. Rice, J. A. Mathematical statistics and data analysis. (Duxbury Press, 1995).

63. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559–575 (2007).

64. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet 381, 1371–1379 (2013).

65. Power, R. A. et al. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. Nat Neurosci 18, 953–955 (2015).

66. Zakharyan, R. et al. Genetic variants of the inflammatory C-reactive protein and schizophrenia in Armenian population: a pilot study. Int. J. Immunogenet. 37, 407–410 (2010).

67. Singh, B. & Chaudhuri, T. K. Role of C-reactive protein in schizophrenia: an overview. Psychiatry Res 216, 277–285 (2014).

68. Wium-Andersen, M. K., Ørsted, D. D. & Nordestgaard, B. G. Elevated C-reactive protein associated with late- and very-late-onset schizophrenia in the general population: a prospective study. Schizophr Bull 40, 1117–1127 (2014).

69. Miller, B. J., Culpepper, N. & Rapaport, M. H. C-Reactive Protein Levels in Schizophrenia. Clin Schizophr Relat Psychoses 1–22 (2013). doi:10.3371/CSRP.MICU.020813

70. Dickerson, F. et al. C-reactive protein is elevated in schizophrenia. Schizophr. Res. 143, 198–202 (2013).

71. Fernandes, B. S. et al. C-reactive protein is increased in schizophrenia but is not altered by antipsychotics: meta-analysis and implications. Molecular Psychiatry (2015). doi:10.1038/mp.2015.87

72. Lewis, C. M. et al. Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. Am. J. Hum. Genet. 73, 34–48 (2003).

73. International Schizophrenia Consortium et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748–752 (2009).

74. Shi, J. et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460, 753–757 (2009).

75. Stefansson, H. et al. Common variants conferring risk of schizophrenia. Nature 460, 744–747 (2009).

76. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. 43, 969–976 (2011).

77. Al-Asmari, A., Al-Asmary, S., Kadasah, S., Arfin, M. & Tariq, M. Genetic variants of interleukin-10 gene promoter are associated with schizophrenia in Saudi patients: A case-control study. North American Journal of Medical Sciences 6, 558 (2014).

78. Saviouk, V., Chow, E. W. C., Bassett, A. S. & Brzustowicz, L. M. Tumor necrosis factor promoter haplotype associated with schizophrenia reveals a linked locus on 1q44. Mol. Psychiatry 10, 375–383 (2005).

79. Hänninen, K. et al. Interleukin-1 beta gene polymorphism and its interactions with neuregulin-1 gene polymorphism are associated with schizophrenia. Eur Arch Psychiatry Clin Neurosci 258, 10–15 (2008).

80. Sekar, A. A natural allelic series of complex structural variants and its influence on the risk of lupus and schizophrenia. (2014).

81. Vaez, A. et al. In Silico Post Genome-Wide Association Studies Analysis of C-Reactive Protein Loci Suggests an Important Role for Interferons. Circ Cardiovasc Genet 8, 487–497 (2015).

82. Shaked, I. et al. Protective autoimmunity: interferon-gamma enables microglia to remove glutamate without evoking inflammatory mediators. J. Neurochem. 92, 997–1009 (2005).

83. Garg, S. K., Banerjee, R. & Kipnis, J. Neuroprotective immunity: T cell-derived glutamate endows astrocytes with a neuroprotective phenotype. J. Immunol. 180, 3866–3873 (2008).

84. Javitt, D. C. Glutamatergic theories of schizophrenia. Isr J Psychiatry Relat Sci 47, 4–16 (2010).

85. Marsman, A. et al. Glutamate in Schizophrenia: A Focused Review and Meta-Analysis of 1H-MRS Studies. Schizophr Bull sbr069 (2011). doi:10.1093/schbul/sbr069

86. Chung, K.-H. et al. The Link between High-Sensitivity C-Reactive Protein and Orbitofrontal Cortex in Euthymic Bipolar Disorder. Neuropsychobiology 68, 168–173 (2013).

87. Hong, E. P., Kim, D. H., Suh, J. G. & Park, J. W. Genetic risk assessment for cardiovascular disease with seven genes associated with plasma C-reactive protein concentrations in Asian populations. Hypertens. Res. 37, 692–698 (2014).

88. Vongpatanasin, W. et al. C-reactive protein causes downregulation of vascular angiotensin subtype 2 receptors and systolic hypertension in mice. Circulation 115, 1020–1028 (2007).

89. Zhou, H. et al. Interleukin 6 augments mechanical strain-induced C-reactive protein synthesis via the stretch-activated channel-nuclear factor κ B signal pathway. Heart 99, 570–576 (2013).

90. Timpson, N. J. et al. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. Lancet 366, 1954–1959 (2005).

91. Davey Smith, G. et al. Association of C-reactive protein with blood pressure and hypertension: life course confounding and mendelian randomization tests of causality. Arterioscler. Thromb. Vasc. Biol. 25, 1051–1056 (2005).

92. Spector, T. D. et al. Low-level increases in serum C-reactive protein are present in early osteoarthritis of the knee and predict progressive disease. Arthritis Rheum. 40, 723–727 (1997).

93. Bos, S. D. et al. Allelic variation at the C-reactive protein gene associates to both hand osteoarthritis severity and serum high sensitive C-reactive protein levels in the GARP study. Ann. Rheum. Dis. 67, 877–879 (2008).

94. Vlad, S. C., Neogi, T., Aliabadi, P., Fontes, J. D. T. & Felson, D. T. No association between markers of inflammation and osteoarthritis of the hands and knees. J. Rheumatol. 38,

1665–1670 (2011).

95. Kerkhof, H. J. M. et al. Serum C reactive protein levels and genetic variation in the CRP gene are not associated with the prevalence, incidence or progression of osteoarthritis independent of body mass index. Ann. Rheum. Dis. 69, 1976–1982 (2010).

96. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium, Hingorani, A. D. & Casas, J. P. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. Lancet 379, 1214–1224 (2012).

97. Anand, S. S. & Yusuf, S. C-reactive protein is a bystander of cardiovascular disease. Eur. Heart J. 31, 2092–2096 (2010).

98. Brunner, E. J. et al. Inflammation, insulin resistance, and diabetes--Mendelian randomization using CRP haplotypes points upstream. PLoS Med. 5, e155 (2008).

99. Timpson, N. J. et al. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. Int J Obes (Lond) 35, 300–308 (2011).

100. Wium-Andersen, M. K., Orsted, D. D. & Nordestgaard, B. G. Elevated C-Reactive Protein, Depression, Somatic Diseases, and All-Cause Mortality: A Mendelian Randomization Study. Biol. Psychiatry (2013). doi:10.1016/j.biopsych.2013.10.009

101. Ryberg, H. et al. Discovery and verification of amyotrophic lateral sclerosis biomarkers by proteomics. Muscle Nerve 42, 104–111 (2010).

102. Kok, E. H. et al. CRP gene variation affects early development of Alzheimer's disease-related plaques. J Neuroinflammation 8, 96 (2011).

103. Song, I.-U., Chung, S.-W., Kim, J.-S. & Lee, K.-S. Association between high-sensitivity C-reactive protein and risk of early idiopathic Parkinson's disease. Neurol. Sci. 32, 31–34 (2011).

104. Pasco, J. A. et al. Association of high-sensitivity C-reactive protein with de novo major depression. Br J Psychiatry 197, 372–377 (2010).

105. Prasad, K. C-reactive protein (CRP)-lowering agents. Cardiovasc Drug Rev 24, 33–50 (2006).

106. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. 14, 507–515 (2013).

107. Vimaleswaran, K. S. et al. Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts. PLoS Med. 10, e1001383 (2013).

108. Smith, G. D. & Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. Int. J. Epidemiol. 33, 30–42 (2004).

109. Jansen, H., Samani, N. J. & Schunkert, H. Mendelian randomization studies in coronary artery disease. European Heart Journal 35, 1917–1924 (2014).

SECTION II:

INTEGRATIVE POST-GWAS ANALYSES AND SYSTEMS GENETICS

8

Genetics of coronary artery disease : Genome-wide association studies and beyond

Atherosclerosis. 2012 Nov;225(1):1-10

Bram P. Prins*, Vasiliki Lagou*, Folkert W. Asselbergs, Harold Snieder, Jingyuan Fu

*equal contribution

ABSTRACT

Genome-wide association (GWA) studies on coronary artery disease (CAD) have been very successful, identifying a total of 32 susceptibility loci so far. Although these loci have provided valuable insights into the etiology of CAD, their cumulative effect explains surprisingly little of the total CAD heritability. In this review, we first highlight and describe the type of genetic variants potentially underlying the missing heritability of CAD: single nucleotide polymorphisms (SNPs) or structural variants, each of which may either be common or rare. Although finding missing heritability is important, we further argue in this review that it constitutes only a first step towards a fuller understanding of the etiology of CAD development. To close the gap between the genotype and phenotype, we propose a systems genetics approach in the post-GWA study era. This approach that integrates genetic, epigenetic, transcriptomic, proteomic, metabolic and intermediate outcome variables has potential to significantly aid the understanding of CAD etiology.

INTERPRETATION AND LIMITATIONS OF GENOME-WIDE ASSOCIATION FINDINGS FOR CAD

Coronary artery disease (CAD) is the leading cause of death in Western societies. For example, in the United States the total prevalence of CAD is 7.0% in adults over 20 years of age and it caused about 1 of every 6 deaths in 2007[1]. It can be viewed to result from a combination of genetic and environmental factors as well as their interactions. Epidemiological studies have identified many traditional risk factors for CAD, including tobacco use, physical inactivity, poor nutrition, obesity, hypertension, high blood cholesterol, and diabetes. In addition to these modifiable risk factors, CAD and its main complication, myocardial infarction (MI), have a strong genetic basis[2]. For example, a family history for CAD was associated with CAD independent of other cardiovascular risk factors[3]. Based on a 36-year follow-up study of more than 20,000 Swedish twins the heritability (h^2) of fatal coronary events was estimated at 0.57 for males and 0.38 for females[4].

The CAD gene database (CADgene)[5] includes information on more than 300 candidate genes, but many of the genetic mechanisms that predispose people to CAD remained unknown until the development of a highly dense genotyping array to analyse common variants genome-wide. This provided new opportunities to identify genetic risk factors associated with CAD. The genome-wide association (GWA) study on CAD was pioneered in 2002 by a Japanese group using a genotyping array of >90,000 single nucleotide polymorphisms (SNPs) in 94 MI cases and 658 controls[6,7]. These early studies pointed to two susceptibility loci (*LTA* and *LGALS2*) that are involved in inflammation and induction of cell-adhesion molecules, but later studies failed to replicate their association[8]. However, a third gene in the same pathway (*BRAP*) yielded one of the strongest associations with a single SNP in a CAD GWA study (OR = 1.42; Table 1)[9,10]. The first wave of seven high-throughput array-based association studies on CAD[9,11–16] identified a total of 12 risk loci with mostly modest effect sizes of odds ratios (ORs) in the 1.1–1.2 range and collectively explaining only a small part of the estimated CAD heritability. For example, the cumulative effect of nine loci identified in the Myocardial Infarction Genetics (MIGen) Consortium explained just 2.8% of the variance in risk for early-onset MI[12]. This suggests that these GWA studies were probably underpowered due to modest sample sizes. Therefore, the majority of CAD heritability remained missing, which limited the clinical translation of

Table 1. Summary of 33 independent risk variants at 31 CAD susceptibility loci identified by GWA studies.

Locus No	Genome region	Risk SNP	Risk allele	AF	Odds ratio	Genes	Association with (traditional) risk factors[a]	Ref
1	1p13.3	rs646776	T	0.81	1.19	*CELSR2, PSRC1, SORT1*	LDL, response to statin, progranulin level, total cholesterol, Lp-PLA2activity and mass	25
2	1p32.2	rs17114036	A	0.91	1.17	*PPAP2B*	—	20
3	1p32.3	rs11206510	T	0.81	1.15	*PCSK9*	LDL	12,20
4	1q41	rs17465637	C	0.72	1.14	*MIA3*	—	12,14,20
5	2q33.1	rs6725887	C	0.14	1.17	*WDR12*	—	12,20
6	3q22.3	rs9818870	T	0.15	1.15	*MRAS*	—	13
7	6p21.31	rs17609940	G	0.75	1.07	*ANKS1A*	—	14
8	6p21.33	rs3869109	G	0.56	1.14	*HCG27, HLA-C*	Triglycerides	24
9	6p24.1	rs12526453	C	0.65	1.12	*PHACTR1*	—	12,20,25
		rs6903956[b]	A	0.03	1.51	*c6orf105*	—	21
10	6q23.2	rs12190287	C	0.62	1.08	*TCF21*	—	20
11	6q25.3	rs3798220	C	0.02	1.92	*SLC22A3, LPAL2, LPA*	Lp(a) level	20
		rs10455872[g]	G	0.07	1.7		—	19
12	7q21	rs1859023[c]	A	0.31	0.72[d]	*PFTK1*	—	26
13	7q22.3	rs10953541	C	0.74	1.08	*BCAP29*	—	25
14	7q32.2	rs11556924	C	0.62	1.09	*ZC3HC1*	—	20
15	9p21.3	rs4977574	G	0.56	1.29	*CDKN2A, CDKN2*[e]	Abdominal aortic aneurysm, intracranial aneurysm	10,12, 14-16,20,25,27
16	9p21.3	rs7865618	A	0.59	1.18	*MTAP*[e]	Type 2 diabetes	27
17	9q34.2	rs579459	C	0.21	1.1	*ABO*	Serum phytosterol level, plasma levels of liver enzymes, venous thromboembolism, E-selectin levels, adhesion levels	20
18	10p11.23	rs2505083	C	0.43	1.08	*KIAA1462*	Non-alcoholic fatty liver, disease histology	20,25
19	10q11.21	rs1746048	C	0.84	1.17	*CXCL12*	—	12,20
20	10q23.2	rs1412444	T	0.37	1.1	*LIPA*	Systolic blood pressure	25,27
21	10q24.32	rs12413409	G	0.89	1.12	*CYP17A1, CNNM2, NT5C2*	Systolic blood pressure, intracranial aneurysm	20,25
22	11q22.3	rs974819	T	0.22	1.07	*PDGFD*	—	25
23	11q23.3	rs964184	G	0.13	1.13	*ZNF259, APOA5-A4-C3-A1*	HDL, hyper-triglyceridemia,	20
24	12q24.12	rs11066001[b]	C	0.34	1.42	*BRAP*[f]	triglycerides	10
	12q24.12	rs671[b]	A	0.23	1.43	*ALDH2*[f]	—	10
25	13q34	rs4773144	G	0.44	1.07	*COL4A1, COL4A2*	—	20
26	14q32.2	rs2895811	C	0.43	1.07	*HHIPL1*	—	20
27	15q25.1	rs3825807	A	0.57	1.08	*ADAMTS7, MORF4L1*[f]	—	20,23
	15q25.1	rs4380028	C	0.65	1.07	*ADAMTS7f*	—	25
28	17p11.2	rs12936587	G	0.56	1.07	*RASD1, SMCR3, PEMT*	—	20
29	17p13.3	rs216172	C	0.37	1.07	*SMG6, SRR*	Aortic root size, type 2 diabetes	20
30	17q21.32	rs46522	T	0.53	1.06	*UBE2Z, GIP, ATP5G1, SNF8*	—	20
31	19p13.2	rs1122608	G	0.75	1.15	*LDLR*	—	12,20
32	21q22.11	rs9982601	T	0.13	1.2	*SLC5A3, MRPS6, KCNE2*	—	12,20

a The associations were extracted from the GWA Catalog database (www.genome.gov/gwastudies/). The traits are listed here if their associated SNPs are in linkage disequilibrium with CAD SNPs (r2 > 0.5, based on the HapMap II and III CEU panel).

b Association detected only in Chinese Han or Japanese populations.

c Assocation detected only in African American populations.

d Hazard Ratio.

e These loci are reported as independent; LD (r2) < 0.3 in the HapMap II and III CEU panel.

f These loci are not reported as independent but map to different genes; r2 > 0.7 for rs11066001 and rs671 in 1000 Genomes Pilot 1 data for CHB + JPT; r2 > 0.5 for
rs3825807 and rs4380028 in the HapMap II CEU panel.

g This variant has been found through a study employing a gene-centric chip designed to assay SNPs in genes implicated in cardiovascular, metabolic and inflammatory disease[28]

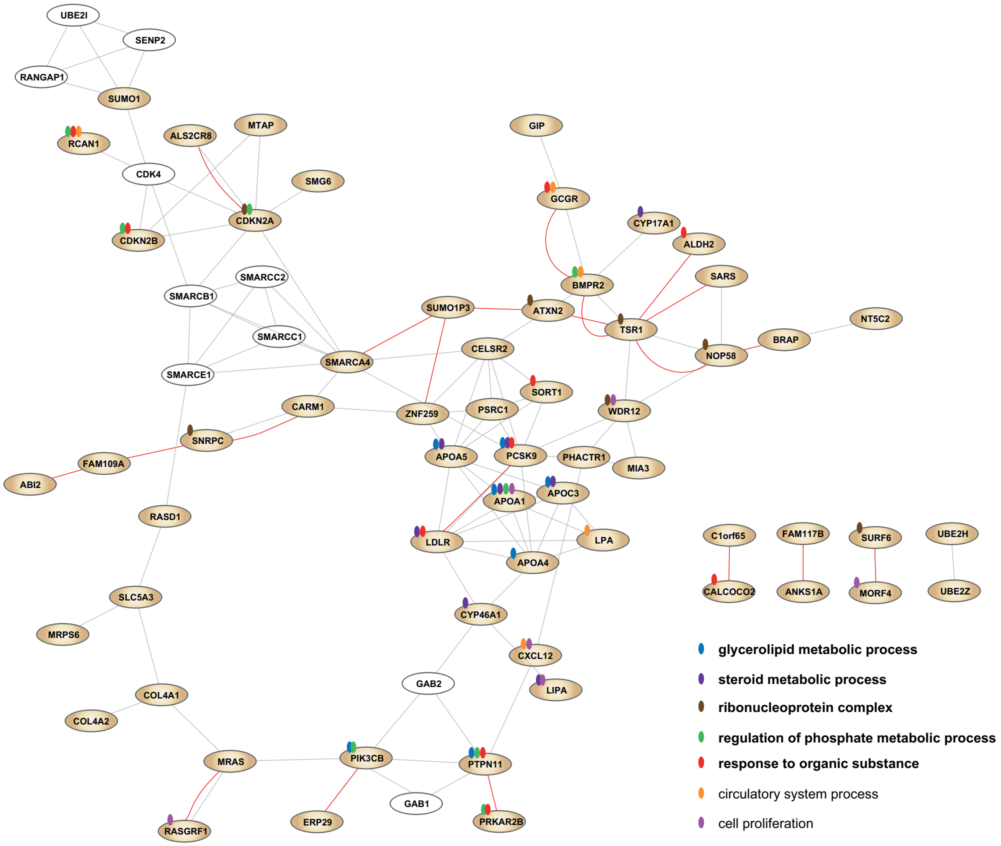


Figure 1. Pathways underlying CAD associated loci. Each colour coded node represents a gene. The genes in brown are candidate genes at CAD associated loci whereas those in white are genes that show functional connections with the CAD genes. The links between genes indicate their functional connection: those in grey are the combined functional connections with median confidence predicted by STRING, those in red are the direct protein–protein interactions predicted by DAPPLE. Selections of enriched biological processes as annotated by Gene ontology are highlighted for each gene. Only the processes in bold remain significant with FDR<0.05 after taking multiple testing into account. Details of the analysis can be found in the supplementary methods.

these genetic findings[17,18]. Realizing this, some of the five most recent GWA studies had very large sample sizes[19–24] (e.g., over 22,000 cases and 64,000 controls in the CARDIoGRAM discovery set) and were conducted on different ethnic groups (Europeans, South Asians, Han Chinese and African Americans). These studies have now brought the number of independent risk variants up to a total of 34 at 32 genomic loci (Table 1).

Most of these loci have small effect sizes with ORs in the 1.05–1.20 range. Interestingly, 12 out of the 32 loci are also associated with traditional CAD risk factors and related traits, including blood pressure, low-density lipoprotein (LDL) cholesterol and plasma lipoprotein(a) [Lp(a)] level. This provides genetic evidence for the causal effect of these traditional risk factors on CAD risk.

Translating GWA signals to biological function is seldom straightforward. Therefore, we conducted a pathway analysis of all the 31 CAD associated loci discovered up until the end of 2011 in order to provide some initial functional insight. A functional connection network and annotation analysis on 86 potential candidate genes underlying the 31 associated loci highlighted several dominant processes. Some of these were expected such as glycerolipid and steroid metabolism, cell proliferation and the circulatory system (Figure 1). However, others may not be immediately obvious and may suggest new hypotheses regarding underlying mechanisms for CAD. Although the recent large-scale GWA studies on CAD have more than doubled the number of risk loci offering more insight into the disease etiology, they did not confirm associations for the majority of the candidate genes from the CADgene database. Moreover, the combined effect of the associated loci still only explains approximately 10% of the additive genetic variance of CAD, leaving the majority of its heritability unexplained[20].

In this review, we first highlight and describe the importance of the potential sources of the missing heritability in CAD/MI. Then we will argue that pinpointing the specific factors underlying the heritability is essential but far from sufficient to fully understand disease etiology for several reasons. First, it will remain a challenge to translate newly identified genetic signals to biological function. Second, numerous important contributors to disease risk are not covered by the heritability estimate, including gene–gene interaction, epigenetic variation and gene–environment interaction. We then propose that a systems biology approach that integrates environmental, genetic and epigenetic factors as well as transcriptomic, proteomic, metabolic and intermediate outcome variables would be the best way forward and would aid significantly to the understanding of CAD etiology.

FACTORS UNDERLYING THE MISSING HERITABILITY

If we limit ourselves to the narrow-sense heritability, which only reflects additive genetic effects (i.e., no dominant or epistatic effects)[29], suggested explanations of missing heritability include additional common SNPs with (very) small effect, rare SNPs with larger effects, and structural variants. Optimal designs, technologies and statistical approaches to detect sources of unexplained heritability for common complex traits and diseases have recently been reviewed in considerable depth[30–33] and leading geneticists have offered their opinions on the subject[34]. Therefore, we discuss these issues only briefly here and focus on the sources of missing heritability and their importance for CAD.

Common SNPs with (very) small effect

Genotyping platforms currently used in GWA studies are designed to tag most known common SNPs (minor allele frequency [MAF] > 0.05), thereby testing the "common disease – common variant" hypothesis. With only a few exceptions, the identified risk alleles of CAD have small to modest effects with ORs between 1.05 and 1.2 and frequencies ranging from 0.13 to 0.91 (Table 1). The as of yet unidentified common risk alleles may have (very) small effects limiting the possibility to detect them individually with the current GWA study sample sizes. This raises the question how many more of these common SNPs with small effect sizes we would need to discover to explain the entire CAD heritability. The total number of such underlying risk variants can be estimated as a function of disease heritability, disease prevalence and some simplifying assumptions that all risk alleles have the same relative risk and frequency[35]. Figure 2 shows the numbers of risk variants for a range of effect sizes (ORs between 1.05 and 1.20) and different allele frequencies for males and females separately. For example, at a disease prevalence of 7%, 1020 and 1218 risk alleles, each with a relative risk of 1.1 and frequency of 0.1, are needed to explain the heritabilities of 0.38 and 0.57 in females and males, respectively.

However, the number of risk alleles increases at an exponential rate with decreasing

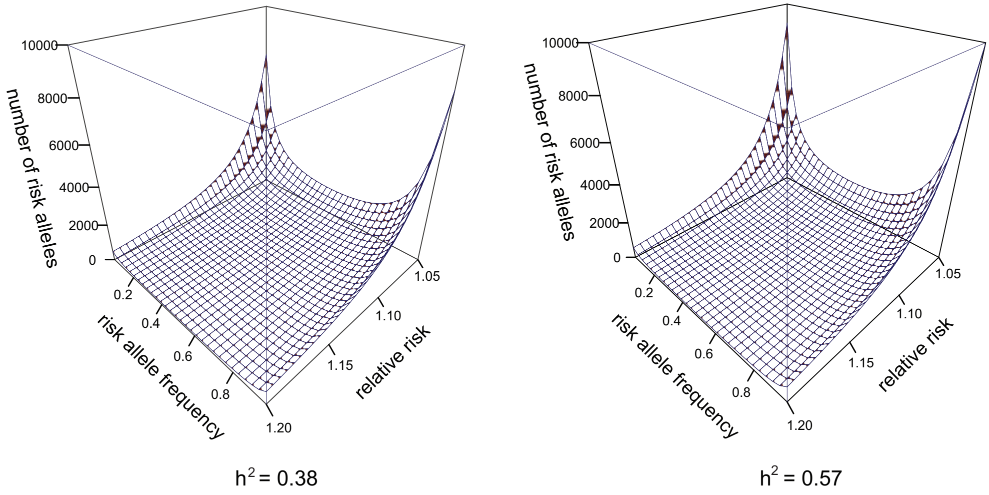


Figure 2. Estimate of the number of common variants that contribute risk to CAD. The number of risk variants underlying heritability can be modelled as $n=\ln[h^2+(1-h^2)K]-\ln(K)\ /\ 2\{\ln[1+p(\lambda^2-1)]-\ln[1+p(\lambda-1)]^2\}$, where n refers to the number of risk variants; h2 is the heritability of the disease; K is the disease prevalence in the population; λ is the relative risk of a risk allele and p is the frequency of the risk allele, assuming all risk alleles have the same relative risk and frequency[35]. At the disease prevalence K of 0.07, the plots show the number of common variants needed to explain heritabilities of 0.38 for females and 0.57 for males based on a 36-year follow-up study of >20,000 Swedish twins[4] depending on the relative risk and allele frequency.

relative risk. If the relative risk decreases to 1.05, the total number of risk alleles increases to 4040 for h2 = 0.38 and 4823 for $h^2 = 0.57$. More sophisticated estimates can be obtained by taking into account the full spectrum of expected risk effects and allele frequencies[36]. The detection of risk alleles with small effect requires exponential increases in sample sizes, because required sample sizes scale approximately quadratically with 1/|(OR-1)|[32]. Realizing this need, international consortia have emerged, such as the Myocardial Infarction Genetics Consortium (MIGen)[12], CARDIoGRAM[20,37] and the Coronary Artery Disease (C4D) genetics consortium[25]. These consortia have performed meta-analyses combining the association signals from multiple GWA studies, maximizing the power to discover risk alleles for CAD.

Rare SNPs with large effect

SNPs with MAF less than 5% in the general population are considered to be of low frequency (i.e., rare). The occurrence of rare variants in the population can be due to selection pressure, random genetic drift or introduction of recent mutations[46,47]. These alleles, although individually rare, are collectively frequent and might contribute substantially to genetic susceptibility underlying complex traits and diseases[31,48]. However, GWA arrays predominantly include common SNPs and in general these do not tag the rare variants well. Therefore our knowledge of the impact of rare variants on CAD remains limited. To address this issue a number of novel experimental strategies and statistical models for the detection of rare variants and their association with complex traits and diseases have emerged[30,49]: 1) as rare variants are often of recent origin they can typically be tagged by the haplotype on which they arose, because recombination has had insufficient time to break down the linkage disequilibrium (LD) surrounding the variant[30,49]; 2) recent advances in high-throughput sequencing technology have accelerated the discovery of rare risk alleles; 3) custom-made arrays have specifically included rare variants in target regions thereby allowing genotyping of such variants in large sample sizes. These advances have uncovered the association between CAD and several rare variants. For example, the association of rs3798220 (MAF: 0.02) at the *SLC2A-LPAL2-LPA* locus was first detected by haplotype association[11] and subsequently identified by a custom-made array[19] (Box 2). Application of haplotype association analysis to the Wellcome Trust Case Control Consortium (WTCCC) GWA data identified rare variants at one known locus (*CDKN2B*) and three novel loci for CAD: *EIF4H*, *HFE2*, and *ZBTB43*[50]. These rare variants often have larger effects than common variants. For example, the OR of rs3798220 is 1.92, much larger than the effects of common variants with an OR between 1.05 and 1.2.

Structural variants

Besides variation at a single nucleotide position, a segment of DNA can be deleted, duplicated or rearranged. This type of DNA variation is known as structural variation. One common type of structural variation is the copy number variant (CNV) that refers to DNA deletion or duplication >1000 base pairs in size[51], which might contribute substantially to risk for common diseases as shown recently for obesity[52]. Previous studies have identified the association of CAD risk with low kringle IV type 2 copy number at the *SLC2A-LPAL2-LPA* locus[44,53] and high number of CA repeats at the *NOS3* locus[54]. Another well-known example in relation to CAD involves Heterozygous Familial Hypercholesterolemia (HeFH), which is an autosomal dominant disorder that affects 1 in 500 people. The genetics underlying the disease in the majority of HeFH patients include SNPs as well as CNVs or small deletions within the LDL receptor gene *LDLR*, making it impossible for the liver to catabolize LDL cholesterol. The resulting rise in plasma LDL cholesterol leads to atherosclerosis and up to a 100 times greater risk of CAD[55].

So far, GWA studies have been unsuccessful in detecting effects of CNVs on CAD, perhaps because they only capture the common CNVs. Despite good coverage of CNVs no significant associations were detected in the MIGen[12] and WTCCC studies[56]. They concluded that common CNVs that can be typed on existing platforms are unlikely to contribute greatly to the genetic basis of common human diseases.

BEYOND THE (NARROW SENSE) HERITABILITY

Box 1. Glossary of terms.

Epigenetic effects – Heritable changes in gene expression that are not caused by changes in DNA sequence, such as DNA methylation (addition of methyl groups to a DNA base) and histone modification (histones are proteins that enable dense packing of DNA in cell nuclei).

Epistasis – Interaction between genes that may result in a phenotype different from the expected phenotype in the case that these genes would not interact.

Heritability – The proportion of individual differences (i.e. variation) in a certain trait (or phenotype) that can be attributed to genetic variation. If the genetic variation includes the total genetic variation, consisting of additive genetic effects, dominance genetic effects (representing interactions between alleles at the same locus), and epistatic genetic effects (representing interactions between alleles at different loci), this is called the broad-sense heritability. If this genetic variation is limited to the additive genetic variation only, this is called the narrow-sense heritability.

Missing heritability – For all of the diseases and traits that have been studied by means of GWA studies, the identified variants explain only a small proportion of the total heritability. The proportion of heritability that remains unaccounted for is generally referred to as the missing heritability.

Haplotype – A set of alleles or variants that is inherited together as a unit.

Linkage disequilibrium (LD) – The extent to which two alleles are non-randomly associated, which is determined by the degree of recombination.

Omics – A suffix that is added to a wide variety of analyses to indicate they occur on a large or genome-wide scale. Transcriptomics for example refers to analysis of genome-wide expression level of messenger RNAs – the transcriptome.

Pathway-based analysis – An approach in which genome-wide results for a trait or disease are analysed and interpreted in the context of predefined pathways, which are collections of genes or proteins with know interaction, instead of investigating the individual effects of genes. This type of analysis is frequently applied as part of post GWAS analyses, to identify potentially important molecular mechanisms underlying the disease or trait of interest.

Post GWAS analyses – A recently coined term that refers to a collection of methods and approaches that aim to reveal the functional consequences of loci identified in GWA studies.

Quantitative trait loci (QTLs) – These are specific regions in the genome that influence a quantitative trait. Examples of quantitative traits include RNA levels (genome-wide referred to as the transcriptome), and levels of metabolites (genome-wide referred to as the metabolome).

Systems genetics – A recently coined term that refers to an integration of genetic analysis approaches aiming to understand the complexity of genotype and phenotype relationships in complex traits and diseases, in analogue fashion to systems biology.

Box 2. Genetic architecture and function of the LPA locus.

Lipoprotein(a) [Lp(a)] levels have long been known to be a risk factor for CAD[38] with very high heritability (~ 90%)[39,40]. It has also been known that this heritability could almost entirely be explained by variation at the apolipoprotein(a) gene on 6q25 as shown in linkage studies[41–43]. The identification of this poster child locus (SLC2A-LPAL2-LPA) as the strongest for CAD to date and elucidation of (part of) its genetic architecture is particularly intriguing. CAD was initially observed to be associated with two haplotypes of four SNPs[11] (rs2048327, rs3127599, rs7767084, and rs10755578) that turned out to tag two rare variants rs3798220 and rs10455872 (see Table 1), and a CNV of kringle IV type 2 (KIV-2) repeats. Furthermore, this locus was observed to be associated with expression of the LPA gene in the liver (rs9355814, $P = 2.24 \text{x} 10^{-28}$)[19]. Interestingly, the CAD associated SNPs rs3798220 and rs10455872 were also highly associated with Lp(a) levels in the serum, together explaining 36% of the total Lp(a) variation. After the adjustment for Lp(a) level, their associations with CAD were abolished, which indicates that Lp(a) level is indeed a causal intermediary factor[19]. Further research showed that the CAD risk variants rs3798220 and rs10455872 together with the KIV-2 repeat explained a larger proportion of variation in Lp(a) concentration than the SNPs by themselves[44]. This suggests that both SNPs and CNVs contribute to CAD risk through their influence on Lp(a) concentrations. A recent GWA study by Kivimäki et al.[45] detected a common SNP (rs783147 with a MAF of 0.47) with a very strong effect on Lp(a) ($P = 3.1 \times 10^{-58}$) that explained 11.7% of its variance. In conclusion, these results show a fairly complicated genetic architecture of the LPA locus with multiple independent variants contributing to Lp(a) levels including rare SNPs, common SNPs and a CNV.

Above we discussed the potential genetic sources of missing (narrow-sense) heritability of CAD. Although important, just finding the missing heritability is only a first step towards a fuller understanding of the disease etiology because the narrow-sense heritability does not capture at least three factors that are believed to be of vital importance for disease development: gene–gene interactions, gene–environment interactions and epigenetic effects.

First, genes do not function in isolation. There is increasing awareness that gene–gene interaction or epistasis plays a role in susceptibility to complex diseases. We observe that disease-associated genes identified by GWA studies often converge on pathways, co-expression networks and protein–protein interaction networks[49] as illustrated by the functional connection network of CAD loci shown in Figure 1. Some gene products even show direct interaction as observed between *PCSK9* and *LDLR*. The proteinase *PCSK9* can bind to the LDL receptor and mediate the degradation of *LDLR*[57]. However, detecting epistatic effects statistically remains challenging. GWA studies have typically used single-locus strategies and a risk variant may thus be missed if its marginal effect is not strong enough to pass the genome-wide significance level. Cordell and others have provided a critical survey of the methods and software to detect interactions in the context of GWA studies and showed that epistasis analysis is statistically feasible[58]. In the near future, pathway-based association analyses are expected to provide a new paradigm for the second-wave of GWA studies[59].

Second, the expression of some genes may be dependent on environmental factors. Sabatti and co-workers performed a GWA analysis of gene–environment interaction for nine metabolic traits in the Northern Finland Birth Cohort[60], including some traditional CAD risk factors such as triglycerides, HDL, LDL, body mass index and blood pressure. Although the gene–environment interactions detected in this study need further replication, it shows that prospectively investigating such interactions for CAD risk may be fruitful. Lanktree and Hegele specifically discussed gene–gene and gene–environment interactions in CAD development and concluded that accounting for gene–gene and gene–environment interactions is important for future strategies of diagnosis, prognosis and management of CAD[44].

Third, one possible mechanism through which environmental factors contribute risk to complex disease such as CAD is through mediation of the epigenome[61]. Epigenetic effects refer to all meiotically and mitotically heritable changes in gene expression that are not coded in the DNA sequence such as those caused by methylation and histone modification. Epigenetic mechanisms collectively enable the cells to respond quickly to environmental changes. Several studies have argued that epigenetic variation is a driving force of development, evolutionary adaption, and complex diseases[62,63]. Recent studies have shown differential global DNA methylation levels in peripheral blood leukocytes in CAD patients compared to controls, but the direction of the effect is inconsistent[61,64–66] due to the limited resolution of the global methylation measures.

INTEGRATION: ROLE OF SYSTEMS GENETICS

Whether part of the heritability or not, integration of abovementioned disparate determinants of disease etiology in a common framework is badly needed. We therefore argue that in the post-GWA era, a systems genetics approach may help us move from finding heritability towards understanding the complex biological networks that underlie complex diseases such as CAD. Systems genetics, by definition, is the approach that studies genetic effects within the larger scope of systems biology, which integrates environmental, genetic and epigenetic factors as well as transcriptomic, proteomic, metabolomic and intermediate (e.g., physiological) outcome variables, ideally within the same population[67].

Variation in methylation states and the abundance of molecular levels (transcripts, proteins and metabolites) in cellular systems can be treated as quantitative traits. Their associated genomic loci are therefore called quantitative trait loci (QTLs) for methylation (methQTL), expression (eQTL), protein (pQTL), metabolites (mQTL), and physiological traits (phQTL), respectively. Studies in humans have investigated the genomic architecture of methylation[68], gene expression[69], lipids[70], and other proteins and metabolites of clinical importance[71]. The resulting comprehensive maps of different QTLs are valuable resources for prioritizing causal variants and designing functional experiments. Integrating such data from multiple molecular levels into explanatory models (i.e., systems genetics) provides a powerful holistic approach to study complex traits and holds several promises.

In the context of evolutionary adaptation, systems genetics may provide insight into the robustness of biological systems and buffering of the propagation of genomic variation to the phenotype level. The HapMap and 1000 Genome projects have catalogued many millions of genetic variants in the human genome. However, robustness at the phenotypic level is essential to keep processes and traits in any living organism within narrow limits, even in the face of all this genetic variation. We were one of the first to provide system-wide molecular evidence for phenotypic buffering using a systems genetics approach in a model system[72]. That is, the largest fraction of genomic and transcriptomic variants is silent at the phenotypic level with only a few influential QTL hot spot regions causing major phenotypic variation across a wide range of environments. These results are in agreement with recent findings that many human diseases share their genetic origin with other diseases to some extent[73]. Fragilities at crucial nodes in the molecular networks may underlie this phenomenon.

One important promise of systems genetics relevant for complex diseases, is its potential to improve understanding of the way genetic information is integrated, coordinated and ultimately transmitted through molecular, cellular and physiological networks to enable higher-order functions and alterations of phenotypes. Causal inference through the construction of causal networks can provide insight into the route from genotype to

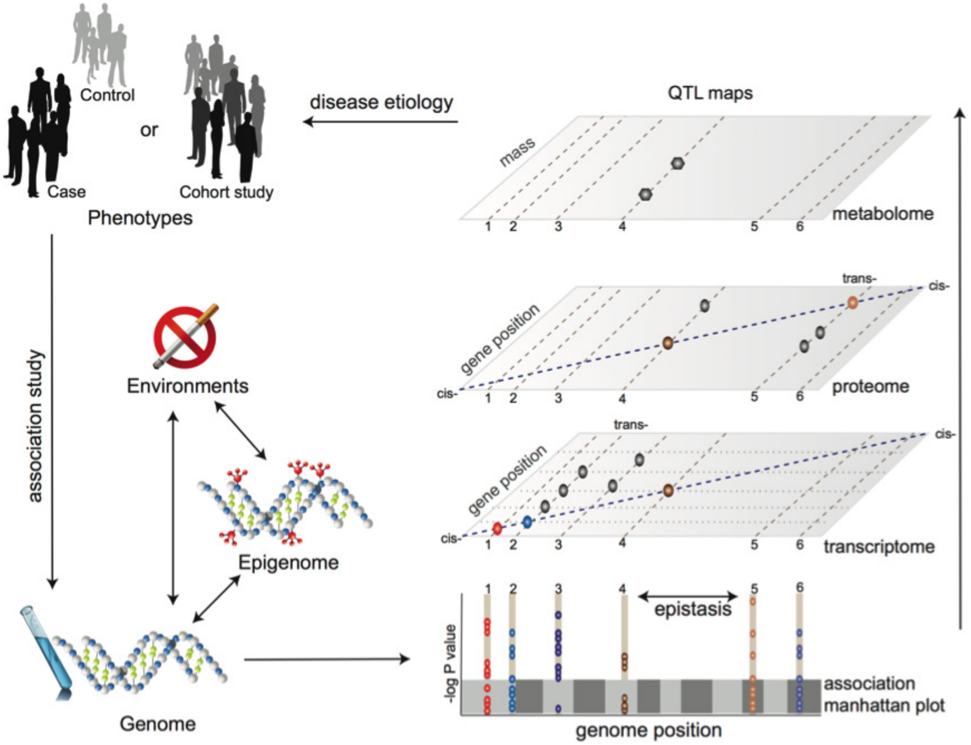


Figure 3. Systems genetics: from finding sources of missing heritability towards understanding the complex biological networks that underlie complex diseases. Systems genetics aims at constructing a holistic view of biological processes by integrating data from multiple molecular levels into explanatory models of complex diseases. Comprehensive "omics" data from transcripts, proteins and metabolites are used in order to explain how these affect the final disease outcome. GWA studies using case–control or cohort designs may discover underlying risk loci as illustrated by the six significant loci in the association ("Manhattan") plot. Above the Manhattan plot, maps of quantitative trait loci (QTLs) at the level of transcriptome (eQTL), proteome (pQTL) and metabolome (mQTL) are shown. The dots on the QTL maps represent the QTLs at the six risk loci. The x-axis of the QTL maps indicates the genome position of the QTLs corresponding to the six risk loci. The y-axis for the eQTL and pQTL maps represents the position of genes affected by the risk variants. If the affected genes physically locate at the risk loci (the dots at the blue dashed diagonal line) these are called cis-QTLs. If the affected genes reside at different genomic regions (the dots at the grey dashed vertical line) these are called trans-QTLs. The y-axis of the mQTL map refers to the mass of the metabolites. Risk variants can have different effects on the different molecular levels. For example, there is only a cis-eQTL at locus1, both cis- and trans-eQTLs at locus 2; a cis-eQTL, a cis- and trans-pQTL and mQTLs at locus 4; cis- and trans-pQTLs at locus 5. Through integration of the genetic effects on multiple levels as well as interactions with environments, the epigenome and amongst the genetic effects (i.e., epistasis), systems genetics endeavours to model the causal network that underlies disease etiology.

phenotype. For example, eQTL maps have provided an important reference source for categorizing the effect of disease-associated SNPs on the expression of genes[74]. SNPs that affect expression of genes at larger distances or on different chromosomes (trans-eQTLs) allow us to identify affected genes downstream, with the potential to reveal novel pathways underlying disease etiology[75]. Similarly, QTLs for proteins and metabolites may coincide with disease-associated SNPs as illustrated in Figure 3 and exemplified by the recent identification of *SORT1* as the causal gene responsible for the CAD GWA signal at the 1p13 locus (Table 1). Munsunuru et al.[76] integrated eQTL and pQTL (for LDL) information at that locus and uncovered that a novel pathway for lipoprotein metabolism regulated by altered expression of the *SORT1* gene underlies the MI etiology.

Systems genetics further offers a means to investigate gene–environment interactions to enhance insight into the pathophysiology of complex diseases. Such interactions involve plasticity of genetic regulation responding to both internal environments (tissues and cell types)[77–80] and external environments[81]. Our recent comparison of gene expression between blood samples and four primary tissues (liver, subcutaneous and visceral adipose tissue and skeletal muscle) characterized four different tissue-dependent manners of genetic regulation of nearby genes (i.e., in cis): specific cis-regulation, alternative cis-regulation, different effect sizes and opposite allelic effects. We further showed that SNPs associated with complex diseases more often exert a tissue-dependent effect on gene expression. As shown for the *SORT1* gene, the MI risk variant alters the expression of *SORT1* in the liver, but not in blood, adipose tissues or muscle[76]. These findings highlight the importance of investigating genetic effects in disease-relevant tissues and environments, in order to correctly characterize the functional effects of disease-associated variants.

CHALLENGES AND PROSPECTS OF SYSTEMS GENETICS FOR CAD

Systems genetics is a powerful method, but applying the approach to the study of complex diseases such as CAD in humans is still a challenge and requires the development of more sophisticated experimental strategies and statistical models. First, the most ideal experiment is to perform system-wide profiling on genome, epigenome, transcriptome, proteome, metabolome and phenome on the same subjects. Integrating “omics” data from different experiments on different subjects can only provide indirect support for etiological hypotheses. Second, we are largely unable to control the effect of environmental factors in human genetics. Environmental factors can have different effects on “omics” levels than on disease endpoints. For example, smoking and diet can have an acute effect on “omics” levels but a chronic effect on disease outcomes. Compared to case–control studies, the prospective cohort study, in which a group of individuals is followed over time and potential disease outcomes predicted on the basis of factors such as genetics, molecular biomarkers, physiological traits and environmental exposures, will become more valuable in human genetic research[82]. The advantages of this cohort design include better definition of environmental exposures and better characterization of disease and risk phenotypes over time. For example, the LifeLines cohort in the Northern three provinces of The Netherlands, will eventually include 165,000 participants that will be followed for 30 years[83]. Approximately 1000 individuals with MIs will be expected in this cohort after five years of follow up. Through integration of systems genetics into the prospective cohort design this study offers great promise for improving our understanding of the causes and prognosis of the burden of CAD. However, considerable investments in bioinformatics and statistical genetics are necessary in order to deliver on this promise, because the complexity of the statistical analysis and required sample size to correctly infer causality constitute a third challenge. Omics data is most valuable when the different layers of data on genome, epigenome, transcriptome, proteome, metabolome and phenome (Figure 3) are mathematically integrated into predictions of the underlying causal networks. However, the robustness of biological systems as mentioned earlier may lead to non-linear relationships between these layers[72]. Even for linear relationships, fairly large sample sizes are required to reliably discriminate between different directions of effects (causal, modifying or independent relationships) between two traits associated with the same locus. A simulation study for this simple scenario showed that a GWA study population size >10,000 is needed to provide 50% sensitivity and 90% positive predictive value for causal inference and realistic QTL effect sizes[84]. On the upside, structural and functional data (gene sequences, gene ontology, metabolic pathways, and protein–protein interactions) as well as independent experimental data gleaned from secondary sources (e.g., gene expression databases) can be used post-hoc to verify the defined gene and causal interferences.

In conclusion, there is no doubt that the GWA approach has been successful in identifying and elucidating previously unexpected genetic candidates for CAD/MI. Including the recent GWA studies with larger sample sizes a total of 32 CAD loci have been identified (Table 1). Despite the numerous successes, the GWA approach has not delivered on some of its promises. A large proportion of heritability remains unexplained and where to find the missing heritability (e.g., largely due to rare variants or to common variants with very small effect) has been hotly debated[32,34]. On the one hand, some studies focused on common variants and estimated that a substantial proportion of variation for a range of common complex traits and diseases can be explained by considering all common SNPs across the genome simultaneously. Examples include Crohn's disease (~24%), bipolar disorder (~41%), type 1 diabetes (~32%), height (~45%), BMI (~17%), von Willebrand factor (~25%) and QT interval (~21%)[85–87]. On the other hand, it was suggested that the GWA associations of common SNPs may result from multiple unobserved rare variants that are in LD with the common SNP; so-called synthetic associations[88]. However, others have argued that the empirical data does not support this hypothesis[89,90]; where both rare and common alleles are uncovered at the same locus, it is much more likely they constitute independent signals[91]. Finally, some studies argued that current estimates of total heritability may be significantly inflated[92], although assumption free methods to estimate heritability do not confirm this[93]. Arguments on the mystery of missing heritability are likely to continue to rage in human genetics and discussions may benefit from complementary information on model organisms such as mouse, rat and Drosophila melanogaster[94]. Here, we further argue that finding (part of) the missing heritability by itself constitutes only a first step towards a fuller understanding of the mechanisms underlying complex diseases. First, complex diseases are the product of the complex interplay between genetic, epigenetic and environmental factors. These interactions are not captured by the (narrow-sense) heritability estimate. Second, even if association can be detected between genotype and phenotype, drawing causal conclusions remains a major challenge. We propose a systems genetics approach within a prospective epidemiological cohort design that integrates molecular traits, including transcripts, metabolites and proteins and a range of (physiological) endophenotypes for CAD. The success of identifying the causal variant and underlying mechanism of the *SORT1* locus illustrates the added value of the systems genetics approach. Although the system-wide application of this approach in humans will require major investments in terms of sample collection, time, and computing-power, in combination with the prospective cohort design it offers great promise in elucidating underlying mechanisms of CAD development. If successful, its findings will not only have implications for disease therapy, but through improvement of risk prediction will also allow prevention efforts to be targeted to those most at risk for CAD[18,95].

Acknowledgements

The authors are indebted to Behrooz Z. Alizadeh and Ilja M. Nolte for their critical review and comments on the manuscript. J. Fu was funded by the Netherlands Organization for Scientific Research (NWO) VENI grant 863.09.007.

Supplementary material

Supplementary Material is available at Atherosclerosis online.

REFERENCES

1. Roger, V. L. et al. Heart disease and stroke statistics--2011 update: a report from the American Heart Association. Circulation 123, e18–e209 (2011).
2. Marenberg, M. E., Risch, N., Berkman, L. F., Floderus, B. & de Faire, U. Genetic susceptibility to death from coronary heart disease in a study of twins. N. Engl. J. Med. 330, 1041–1046 (1994).
3. Schildkraut, J. M., Myers, R. H., Cupples, L. A., Kiely, D. K. & Kannel, W. B. Coronary risk associated with age and sex of parental heart disease in the Framingham Study. Am. J. Cardiol. 64, 555–559 (1989).
4. Zdravkovic, S. et al. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. J. Intern. Med. 252, 247–254 (2002).
5. Liu, H. et al. CADgene: a comprehensive database for coronary artery disease genes. Nucleic Acids Res. 39, D991–996 (2011).
6. Ozaki, K. et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat. Genet. 32, 650–654 (2002).
7. Ozaki, K. et al. Functional variation in LGALS2 confers risk of myocardial infarction and regulates lymphotoxin-alpha secretion in vitro. Nature 429, 72–75 (2004).
8. Kimura, A. et al. Lack of association between LTA and LGALS2 polymorphisms and myocardial infarction in Japanese and Korean populations. Tissue Antigens 69, 265–269 (2007).
9. Ozaki, K. et al. SNPs in BRAP associated with risk of myocardial infarction in Asian populations. Nat. Genet. 41, 329–333 (2009).
10. Takeuchi, F. et al. Genome-wide association study of coronary artery disease in the Japanese. Eur. J. Hum. Genet. 20, 333–340 (2012).
11. Trégouët, D.-A. et al. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat. Genet. 41, 283–285 (2009).
12. Myocardial Infarction Genetics Consortium et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat. Genet. 41, 334–341 (2009).
13. Erdmann, J. et al. New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat. Genet. 41, 280–282 (2009).
14. Samani, N. J. et al. Genomewide association analysis of coronary artery disease. N. Engl. J. Med. 357, 443–453 (2007).
15. McPherson, R. et al. A common allele on chromosome 9 associated with coronary heart disease. Science 316, 1488–1491 (2007).
16. Helgadottir, A. et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316, 1491–1493 (2007).
17. Ioannidis, J. P. A. Prediction of cardiovascular disease outcomes and established cardiovascular risk factors by genome-wide association markers. Circ Cardiovasc Genet 2, 7–15 (2009).
18. Humphries, S. E., Drenos, F., Ken-Dror, G. & Talmud, P. J. Coronary heart disease risk prediction in the era of genome-wide association studies: current status and what the future holds. Circulation 121, 2235–2248 (2010).
19. Clarke, R. et al. Genetic variants associated

with Lp(a) lipoprotein level and coronary disease. N. Engl. J. Med. 361, 2518–2528 (2009).

20. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet. 43, 333–338 (2011).

21. Wang, F. et al. Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population. Nat. Genet. 43, 345–349 (2011).

22. Erdmann, J. et al. Genome-wide association study identifies a new locus for coronary artery disease on chromosome 10p11.23. Eur. Heart J. 32, 158–168 (2011).

23. Reilly, M. P. et al. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. Lancet 377, 383–392 (2011).

24. Davies, R. W. et al. A genome-wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility complex. Circ Cardiovasc Genet 5, 217–225 (2012).

25. Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat. Genet. 43, 339–344 (2011).

26. Barbalic, M. et al. Genome-wide association analysis of incident coronary heart disease (CHD) in African Americans: a short report. PLoS Genet. 7, e1002199 (2011).

27. Wild, P. S. et al. A genome-wide association study identifies LIPA as a susceptibility gene for coronary artery disease. Circ Cardiovasc Genet 4, 403–412 (2011).

28. Keating, B. J. et al. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS ONE 3, e3583 (2008).

29. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. Nat. Rev. Genet. 9, 255–266 (2008).

30. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. 11, 773–785 (2010).

31. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. 11, 415–425 (2010).

32. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).

33. McCarroll, S. A. Extending genome-wide association studies to copy-number variation. Hum. Mol. Genet. 17, R135–142 (2008).

34. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11, 446–450 (2010).

35. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 17, 1520–1528 (2007).

36. Park, J.-H. et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. 42, 570–575 (2010).

37. Preuss, M. et al. Design of the Coronary

ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. Circ Cardiovasc Genet 3, 475–483 (2010).

38. Rhoads, G. G., Dahlen, G., Berg, K., Morton, N. E. & Dannenberg, A. L. Lp(a) lipoprotein as a risk factor for myocardial infarction. JAMA 256, 2540–2544 (1986).

39. Snieder, H., van Doornen, L. J. & Boomsma, D. I. The age dependency of gene expression for plasma lipids, lipoproteins, and apolipoproteins. Am. J. Hum. Genet. 60, 638–650 (1997).

40. Snieder, H., van Doornen, L. J. & Boomsma, D. I. Dissecting the genetic architecture of lipids, lipoproteins, and apolipoproteins: lessons from twin studies. Arterioscler. Thromb. Vasc. Biol. 19, 2826–2834 (1999).

41. Beekman, M. et al. Two-locus linkage analysis applied to putative quantitative trait loci for lipoprotein(a) levels. Twin Res 6, 322–324 (2003).

42. Boerwinkle, E. et al. Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. J. Clin. Invest. 90, 52–60 (1992).

43. Falchi, M. et al. Identification of QTLs for serum lipid levels in a female sib-pair cohort: a novel application to improve the power of two-locus linkage analysis. Hum. Mol. Genet. 14, 2971–2979 (2005).

44. Lanktree, M. B. & Hegele, R. A. Gene-gene and gene-environment interactions: new insights into the prevention, detection and management of coronary artery disease. Genome Med 1, 28 (2009).

45. Kivimäki, M. et al. Conventional and Mendelian randomization analyses suggest no association between lipoprotein(a) and early atherosclerosis: the Young Finns Study. Int J Epidemiol 40, 470–478 (2011).

46. Iyengar, S. K. & Elston, R. C. The genetic basis of complex traits: rare variants or 'common gene, common disease'? Methods Mol. Biol. 376, 71–84 (2007).

47. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. 69, 124–137 (2001).

48. Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. & Amos, C. I. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am. J. Hum. Genet. 82, 100–112 (2008).

49. Lage, K. et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. 25, 309–316 (2007).

50. Zhu, X., Feng, T., Li, Y., Lu, Q. & Elston, R. C. Detecting rare variants for complex traits using family and unrelated data. Genet. Epidemiol. 34, 171–187 (2010).

51. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. Science 305, 525–528 (2004).

52. Bochukova, E. G. et al. Large, rare chromosomal deletions associated with severe early-onset obesity. Nature 463, 666–670 (2010).

53. Kraft, H. G. et al. Apolipoprotein(a) kringle IV repeat number predicts risk for coronary heart disease. Arterioscler. Thromb. Vasc. Biol. 16, 713–719 (1996).

54. Stangl, K. et al. High CA repeat numbers in intron 13 of the endothelial nitric oxide synthase gene and increased risk of coronary artery disease. Pharmacogenetics 10, 133–140 (2000).

55. Yuan, G., Wang, J. & Hegele, R. A. Heterozygous familial hypercholesterolemia: an underrecognized cause of early cardiovascular disease. CMAJ 174, 1124–

1129 (2006).

56. Wellcome Trust Case Control Consortium et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713–720 (2010).

57. Kwon, H. J., Lagace, T. A., McNutt, M. C., Horton, J. D. & Deisenhofer, J. Molecular basis for LDL receptor recognition by PCSK9. Proc. Natl. Acad. Sci. U.S.A. 105, 1820–1825 (2008).

58. Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. Nat. Rev. Genet. 10, 392–404 (2009).

59. Peng, G. et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur. J. Hum. Genet. 18, 111–117 (2010).

60. Sabatti, C. et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat. Genet. 41, 35–46 (2009).

61. Baccarelli, A. et al. Repetitive element DNA methylation and circulating endothelial and inflammation markers in the VA normative aging study. Epigenetics 5, 222–228 (2010).

62. Feinberg, A. P. & Irizarry, R. A. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc. Natl. Acad. Sci. U.S.A. 107 Suppl 1, 1757–1764 (2010).

63. Kong, A. et al. Parental origin of sequence variants associated with complex diseases. Nature 462, 868–874 (2009).

64. Castro, R. et al. Increased homocysteine and S-adenosylhomocysteine concentrations and DNA hypomethylation in vascular disease. Clin. Chem. 49, 1292–1296 (2003).

65. Kim, M. et al. DNA methylation as a biomarker for cardiovascular disease risk. PLoS ONE 5, e9692 (2010).

66. Sharma, P. et al. Detection of altered global DNA methylation in coronary artery disease patients. DNA Cell Biol. 27, 357–365 (2008).

67. Nadeau, J. H. & Dudley, A. M. Genetics. Systems genetics. Science 331, 1015–1016 (2011).

68. Morris, M. R. et al. Genome-wide methylation analysis identifies epigenetically inactivated candidate tumour suppressor genes in renal cell carcinoma. Oncogene 30, 1390–1401 (2011).

69. Staal, F. J. T. et al. Genome-wide expression analysis of paired diagnosis-relapse samples in ALL indicates involvement of pathways related to DNA replication, cell cycle and DNA repair, independent of immune phenotype. Leukemia 24, 491–499 (2010).

70. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010).

71. Illig, T. et al. A genome-wide perspective of genetic variation in human metabolism. Nat. Genet. 42, 137–141 (2010).

72. Fu, J. et al. System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nat. Genet. 41, 166–167 (2009).

73. Zhernakova, A., van Diemen, C. C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat. Rev. Genet. 10, 43–55 (2009).

74. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. Nat. Rev. Genet. 10, 184–194 (2009).

75. Fehrmann, R. S. N. et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype

converge on intermediate genes, with a major role for the HLA. PLoS Genet. 7, e1002197 (2011)

76. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466, 714–719 (2010).

77. Dimas, A. S. et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325, 1246–1250 (2009).

78. Gerrits, A. et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. PLoS Genet. 5, e1000692 (2009).

79. Nica, A. C. et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 7, e1002003 (2011).

80. Fu, J. et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 8, e1002431 (2012).

81. Li, Y. et al. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet. 2, e222 (2006).

82. Manolio, T. A., Bailey-Wilson, J. E. & Collins, F. S. Genes, environment and the value of prospective cohort studies. Nat. Rev. Genet. 7, 812–820 (2006).

83. Stolk, R. P. et al. Universal risk factors for multifactorial diseases: LifeLines: a three-generation population-based study. Eur. J. Epidemiol. 23, 67–74 (2008).

84. Li, Y., Tesson, B. M., Churchill, G. A. & Jansen, R. C. Critical reasoning on causal inference in genome-wide linkage and association studies. Trends Genet. 26, 493–498 (2010).

85. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. 88, 294–305 (2011).

86. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569 (2010).

87. Yang, J. et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 43, 519–525 (2011).

88. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. PLoS Biol. 8, e1000294 (2010).

89. Anderson, C. A., Soranzo, N., Zeggini, E. & Barrett, J. C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol. 9, e1000580 (2011).

90. Wray, N. R., Purcell, S. M. & Visscher, P. M. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. 9, e1000579 (2011).

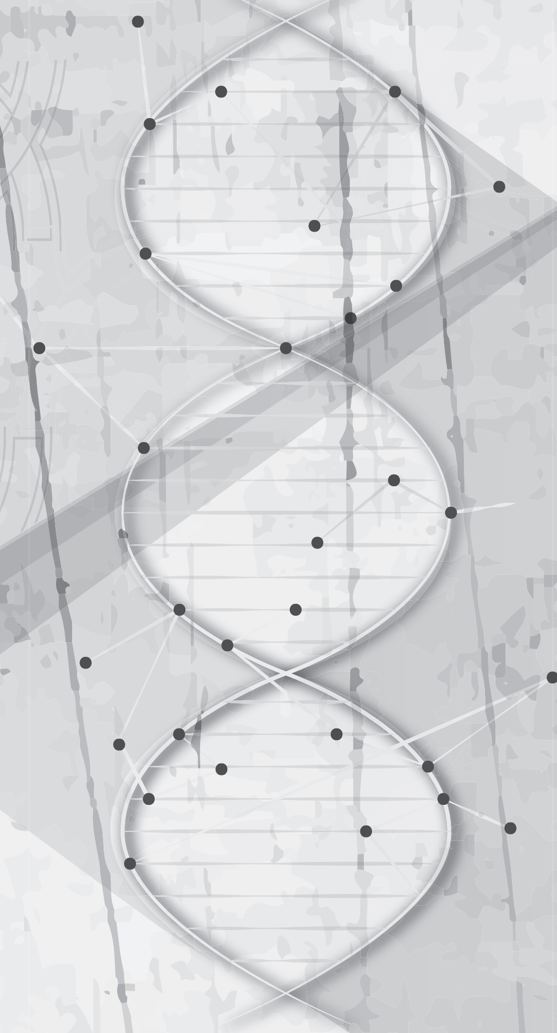
91. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. Am J Hum Genet 90, 7–24 (2012).

92. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. PNAS 109, 1193–1198 (2012).

93. Visscher, P. M. et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2, e41 (2006).

94. Aitman, T. J. et al. The future of model organisms in human disease research. Nat. Rev. Genet. 12, 575–582 (2011).

95. Ashley, E. A. et al. Clinical assessment incorporating a personal genome. Lancet 375, 1525–1535 (2010).



SECTION III : GENERAL SYNTHESIS, BROADER AND FUTURE PERSPECTIVES

9

Discussion & Future Perspectives

DISCUSSION

Section I : GWAS and inflammatory marker genetics

In the first section of this thesis I focused on elucidating the genetic basis for a set of clinically relevant inflammatory markers. Following developments in DNA chip genotyping, both upscaling and miniaturisation, the analysis of many genetic variants in one experiment in a feasible timeframe and for reasonable costs became possible and resulted in the first GWAS experiments[1,2].

For inflammatory markers, the first GWAS was performed in 2007 for C-Reactive Protein (CRP), Interleukin 6 (IL6), and Tumor Necrosis Factor (TNF) as part of a systematic GWAS on biomarkers measured in up to 1008 individuals, where only for CRP one genome-wide significant association was identified[3]. In later years, the power to detect additional loci improved, by ever-increasing sample sizes and improved granularity by means of imputation to reference panels, resulting in the identification of additional loci for each trait, but still leaving a large part of the heritability for these traits unexplained. When assessing prior GWAS work up to now, several important general observations are evident:

I) Firstly the GWAS paradigm is a powerful approach to identify genetic variation underlying complex traits; a brief glance at the GWAS catalog[4] shows that from 2009, the number of SNPs found associated with complex traits ran into the 1000s, previously considered impossible through either candidate gene association or linkage studies.

II) Secondly, the common disease – common variants hypothesis appeared to be correct; for many diseases and traits it appears that many different genetic common variants throughout the genome contribute to the architecture of complex traits, albeit with small effects.

III) Even though many genetic loci have been identified with small effects a large portion of the heritability remains unexplained for the majority of traits.

IV) Genetic variants involved in certain traits can be population specific, and pinpointing causal variants remains difficult.

V) Lastly, stringent quality control for GWAS results is essential to avoid false positives.

Learning from these observations, it is to this end, that in section I of this thesis I aimed to further elucidate the genetic background of some of the clinically most relevant inflammatory biomarkers, namely IL6, TNF, total protein and albumin, by substantially increasing sample sizes, in some cases including data from different populations and narrowing down genomic locations containing causal variants, all supported by the development of a standardised GWAS quality control pipeline.

The need for high-quality GWAS results and high-throughput QC

In Chapter 2, we first present a software suite that enables automated quality control for GWAS summary statistics[5]. The output from GWAS typically is a plain text file, having several columns containing summary statistics, such as p-values and effect sizes and information about the genomic location and quality of a variant. When contributing GWAS results to a meta-analysis consortium, the summary statistic files generated by analysis programs are normally first pre-processed with custom-written scripts so as to meet the standard format outlined by the relevant GWAS consortium and as required by meta-analysis software. As in a typical GWAS 2.5 to 3 million variants are analysed, the resulting files are several millions of lines in length and multiple columns in width, which does not allow manual checking, especially since one sometimes has to deal with hundreds of files for a consortium. However, failing to perform quality control can have various negative consequences, from files being non-processable, to identifying false-positive hits and publishing these[6]. We therefore developed an R-package that could automatically process GWAS result files, visualizing various quality parameters at once, and simultaneously formatting the output results in a desired format. At the time of development, to our knowledge only one other pipeline had been developed to semi-automatically check GWAS results, albeit with several crucial drawbacks, the most important being the lack of: cross-study comparable QC log file generation, the ability to generate clean standardised GWAS output files, the automatic serial throughput of files without writing custom scripts, and the harmonisation of variant definitions using a reference panel of SNPs, all of which are currently implemented in our QCGWAS R-package. This package has been used to perform the QC for the GWAS results in Chapters 3 to 5 and has saved considerable time and efforts.

Nevertheless, there remains room for improvement. Firstly, since imputation reference panels become increasingly large, the time to process result files, as well as memory requirements increase accordingly. Whereas Hapmap-based result files as in Chapters 3 to 5 typically contain 2.5 to 3 million variants, taking between 5 and 15 minutes per file on an ordinary desktop PC, with a memory usage between 2 and 3 GB, more recent imputation panels such as the 1000 Genomes reference based data with over 10 million variants, take over 40 minutes and 20 GB of RAM, far too much when having to process hundreds of files. Improvements on processing speed can be sought in multiple ways ; by for example writing parts of the pipeline in lower-level programming languages such as C++, which can speed up certain computations several hundred fold[7]. Additionally, when keeping R architecture, various packages have been developed to scale-up R capabilities to handle 'big data', by giving R parallel computing capabilities[8], and resolving it's memory devouring habits, such as the *ff* and *bigmemory* packages[9,10] that cleverly make use of the hard disk as memory space and memory-mapping tools. A second major improvement for QCGWAS could be the automatic checking of GWAS results against known positive controls for a certain trait. Now that many thousands of variants are identified for hundreds of traits, positive controls can be easily (and automatically) retrieved from the GWAS catalog, enabling effect-size consistency checks for individual GWAS results, allowing the identification of a multitude of problems, such as whether the correct trait was analysed, or whether perhaps sample mix-ups have occurred and so on. A third and last idea for improvement can be the variant strand alignment checking to a reference for ambiguous SNPs. Currently, ambiguous SNPs whose allele-frequency differ strongly from that of the reference are marked as suspect (i.e. not properly aligning, in QCGWAS defined as < 0.35. versus > 0.65 allele frequency difference compared to the reference), though the frequency of variants can obviously differ between studies and populations, and proper alignment by looking at frequencies is very hard to determine if the minor allele frequency is nearing 50%. Instead, use can be made of linkage disequilibrium patterns in the reference data. Proper strand assignment for ambiguous SNPs can potentially be achieved by comparison of these to frequencies of nearby non-ambiguous SNPs that are in very high LD with the ambiguous SNP, that tend to have very similar allele frequencies.

Biological insights from large-scale meta-GWAS analyses for 3 major inflammatory markers

In Chapters 3 to 5 we performed the largest-ever meta-analyses for four major biomarkers associated with inflammation, being TNF, and IL6, total protein and albumin concentrations. The last two are used for various clinical purposes.

For TNF, we were the first to identify reliable genomic loci, 3 in total, associated with its circulating levels in blood. This is a classic example of the merits of pooling GWAS results so as to increase statistical power to detect loci. Our 3 loci for TNF provide interesting biological insights. Reassuringly, we, for example, showed that two of our loci harbour genes for cytokines and regulators of cytokines (6p21 / LTA and 12q24 / *SH2B3*). Interestingly, our third locus provided evidence for involvement in lipid metabolism (15q21 / *LIPC*). It has been long known that TNF is a very potent lipid metabolism regulator, but our study for the first time pinpoints a genetic locus that may provide a link to this involvement.

Similarly, for IL6 we were able to report 3 loci, two of which were novel. Up to now, only the well-established IL6R locus has been known for its involvement in determining IL6 levels in blood, which is a highly pleiotropic locus harbouring variants influencing various traits including CRP, and fibrinogen, but also diseases such as Coronary Artery Disease (CAD), Rheumatoid Arthritis (RA) and asthma. The two novel loci, 2q14 and 6p21 harbour well-known inflammation related genes, (*IL1F10* and *IL1RN* for 2q14 and *HLA-DRB1/ DRB5* for 6p21.

In the meta-analysis of serum-albumin and total protein concentration, we identified six loci for serum albumin that harbour genes related to ribosome function (19q13 / *HPN-SCN1B*, 2p23 / *GCKR-FNDC4*, 17p13 / *SERPINF2-WDR81*, 18q21 / *TNFRSF11A-ZCCHC2*, 15q15 / *FRMD5-WDR76*, and 19q13 *RPS11-FCGRT*), and 3 loci associated with serum total protein (17p11 / *TNFRS13B,* 6q21.3, and 5q15 / *ELL2*), harbouring genes related to immune function[11]. In this case, we demonstrated that lead SNPs at the identified loci have themselves been reported or are in moderate to high LD (r2 > 0.5) with those reported in the GWAS catalogue for a very diverse range of human complex traits, and reassuringly, are enriched for metabolic phenotypes that are associated with, or are direct products of, protein metabolism.

Fine-mapping

The majority of GWAS lead SNPs are intergenic, and even in the case they do reside within a gene, their functional impact remains difficult to assess when not directly affecting protein-coding (i.e. being intronic). In addition, pin-pointing the variants within a locus that make an actual causal contribution to the trait of interest remains challenging, as it is not necessarily the top-signal within a locus that might be causally contributing. Therefore, in Chapter 5, we applied two subsequent methods to substantially improve fine-mapping of genomic loci involved in determining serum levels of total protein and albumin[11]. First, we applied a novel GWAS meta-analysis method, MANTRA, that facilitates further narrowing down a genomic region containing actual causal variants. It allows for heterogeneity in allelic effects between ancestry groups using a Bayesian approach. This method capitalises on the varying LD structures between populations of different ancestry, which enabled us to define narrower regions of association. Secondly, we defined ''credible sets'' of SNPs, contained in a genomic window that with 99.9% probability contains the causal variants, which are those variants with the strongest signals of association and, hence, most likely to be causal (or tagging an unobserved causal variant). In total, we identified nine novel loci associated with the traits investigated. We observed improved resolution, in terms of the number of SNPs and the genomic interval covered by the credible set for various loci for both traits. The most striking improvements in resolution were observed at the 6q21.3 locus for total protein, wherein the 99% credible set was reduced from 14 SNPs (covering 346 kB), to just three (covering 37 kB). Furthermore, after trans-ethnic meta-analysis, the posterior probability that the lead SNP was causal (or tagged an unobserved causal variant) was more than 95% at 2p23 / *GCKR-FNDC4* and 17p13 / *SERPINF2-WDR81* for serum albumin and at 17p11 / *TNFRS13B* and the 6q21.3 locus for total protein. Given the results of these efforts, fine-mapping using a combination of methods should be a standard component of each (meta-) GWAS analysis.

Functional impact

Traditionally, lead variants are typically annotated using databases such as ANNOVAR[12] as done in Chapters 3 to 5, to obtain initial biological insights as whether these are protein coding or not and if so whether nucleotide substitutions result in amino-acids changes that have functional consequences. One way to assess functional impact of variants on a phenotypic level is to correlate the genotypes of lead variants with the mRNA expression of nearby genes, referred to as (cis-) expression quantitative trait locus (eQTL) mapping. In Chapter 5, we mapped lead SNPs at four of the identified loci to cis expression levels of 18 genes, for which one of the strongest associations was observed for expression of *HLA- DQA1/2,* which is a human leukocyte antigen (HLA) class II antigen with an immune system role related to processing and presentation of antigen peptides. Even though this way initial clues about functional impact of variants can be elucidated, higher-order effects on protein and phenotypical levels cannot be resolved. Instead, a direct way to test effects of variants or genes they reside in can be done through the use of model animals such as zebrafish or mice. For example, in Chapter 5 we identified an association in the 19q13 / *HPN-SCN1B* locus, at which the lead SNP is an intronic SNP within *HPN*, a gene encoding hepsin. Hepsin is a membrane-bound serine protease with substrate specificity for basic amino acids similar to those involved in proalbumin processing, which suggests a physiologic role of hepsin in the cleavage of proalbumin to albumin. By knocking out hepsin in a mouse model, and comparing serum protein and albumin concentrations between these knockout mice and wild-type litter mates, we found overwhelming evidence of reduced serum albumin and, to a lesser extent, reduced total protein, showing that the use of animal models can be a valuable tool to provide functional evidence supporting GWAS findings.

Section II : Integrative post-GWAS analyses and aetiological involvement

Having identified and fine-mapped genetic associations is the first step towards understanding the molecular basis of traits. In order to understand how a phenotype is established through changes in DNA sequence, the functional impact of variants must first be understood, both individually, but also in combination on a systemic level, as I have pursued in Section II of this thesis.

Systems genetics

Even when we have established functional consequences on mRNA and protein levels, the exact molecular mechanisms through which variants and the genomic loci they reside in exert their effects on the phenotype cannot be resolved by evaluating these individually. Instead, clues about how these act in concert can be gathered using various approaches, collectively known as systems genetics. In Chapter 5, we used two complementary approaches: we performed a pathway analysis using MAGENTA[13], which is a method that evaluates the enrichment of sets of genes harbouring our meta-analysis variants that belong to pre-defined biological pathways. Amongst identified pathways we identified RNA processing and protein-trafficking-related pathways. Secondly, one can also look at how the proteins encoded by genes mapping to our loci interact. Proteins that heavily interact are likely to belong to a confined part of a molecular pathway and

/ or physiological process. Using programs such as Cytoscape[14], that can model such interactions, together with clustering algorithms and pathway annotation, we identified several clusters of strongly interacting proteins. These clusters were enriched for ribosomal functioning and protein translation, proteasomal protein degradation, and immune response signalling. Both the pathway analysis and protein-interaction analyses identified molecular processes that were most relevant to the investigated phenotypes. On the basis of the functional analysis approaches taken in Chapter 5, collectively referred to as 'post-GWAS' analyses, and following suggestions in Chapter 8[15], we developed a standardised pipeline in Chapter 6 that integrates a number of in-silico functional approaches[16]. In brief, in Chapter 6 we first performed 'in silico' sequencing, exploring the vicinity of GWAS SNPs to identify all linked variants. In the second phase, we performed eQTL analyses, where we attempted to identify all nearby genes whose expression levels are associated with the corresponding GWAS SNPs. These two phases generated a number of relevant genes that served as input to the next phase of functional network analysis using GeneMANIA[17,18] and Cytoscape[14]. The application of this pipeline to loci identified in the at that time largest ever meta-GWAS analysis done for CRP, yielded a range of enriched biological processes such as the acute-phase response or the acute inflammatory response. About one third of the significantly enriched terms were related to immunologic processes, cytokines, and interferons. Interestingly, the majority of the identified processes pointed to an involvement of interferon biology, in particular, type I interferon associated biological processes, of which the connection to CRP was previously not widely recognised.

Clinical relevance and causal involvement

In Chapters 3 to 5 we identified genomic loci or in some cases specific variants for the various biomarkers that affect a wide range of diseases, including common cardiovascular traits and diseases such as coronary heart disease and diastolic and systolic blood pressure, but also auto-immune diseases such as rheumatoid arthritis, type 1 diabetes and celiac disease. In other words, these loci appear to be pleiotropic. However, identifying genetic overlap between biomarkers and disease does not necessarily imply causal involvement, even though various biomarkers investigated in this thesis are correlated with disease status, used in clinical practice to evaluate disease progression and prognosis, or in the case of TNF and IL6, even used as direct targets of intervention.

Amongst the biomarkers investigated in this thesis, CRP has long been suspected to be causally involved in a number of diseases. Epidemiological studies have shown that CRP is associated with and exhibited a reliable predictive value for cardiovascular disease[19,20], type 2 diabetes[21], and immunity-related disorders such as inflammatory bowel disease (IBD)[22], rheumatoid arthritis[23] and all-cause mortality[24]. Nevertheless, the evidence for a causal involvement of CRP from traditional experimental or observational studies remains controversial. Causal involvement can however be evaluated by means of an approach known as Mendelian randomization (MR). These types of analyses make use of genetically informed instrumental variables (IVs), that is, sets of genetic variants known to have an effect on the exposure of interest, to actually model c.q. genetically reflect this exposure, in this case CRP levels.

In Chapter 7 we selected a panel of 32 different traits and diseases in five broad classes, (i.e. auto-immune-inflammatory, cardiovascular, metabolic, neuro-degenerative and psychiatric), for which association with elevated CRP levels in epidemiological studies has been observed, and performed MR analyses genetically modelling levels of CRP to test causal effects of these on the selected traits. For our MR analyses, we constructed two genetic IV's, one using four SNPs representing only the CRP gene, to avoid any pleiotropic effects - one of the assumptions necessary for performing MR, and another IV using up to 18 SNPs that were associated with CRP levels at genome-wide significance in the largest GWAS for CRP to date. We identified a significant causal protective association of CRP and schizophrenia, and nominal, yet to be confirmed, evidence for a causal involvement in psoriatic osteoarthritis, rheumatoid arthritis, knee osteoarthritis, SBP, DBP, serum albumin levels, and bipolar disorder. We could however not demonstrate any causal effect of CRP on other common somatic and neuropsychiatric outcomes investigated in this study. The apparent non-causal involvement of CRP in the other sets of traits also means that drug interventions targeting CRP for these traits will unlikely be of benefit.

An important aspect of this study was not only the identification of the causal association between CRP and schizophrenia, but the mere fact that individual-level genotype data was not required for the analyses, and instead we could make use of publicly available GWAS summary statistics, which contributed substantially to the feasibility of this study. Nevertheless, these types of MR approaches have to be carefully applied. MR analyses require several assumptions, the violation of which can introduce severe bias. One of the most precarious is that the instruments need to be "pleiotropy-free", i.e., the genetic instruments used in the MR analysis may not be associated with any unmeasured phenotype that is related to the outcome of interest. We identified several genes being pleiotropic through assessing heterogeneity of effects of variants, some of which with well-known pleiotropic effects, such as *GCKR* (mapped in our study by rs1260326), absence of heterogeneity of effects does not automatically imply absence of pleiotropy. Though absence of pleiotropy is rather difficult to be proven empirically, there are several additional tests that can be utilised to try to falsify assumptions about pleiotropy and thus to minimize the chance of bias[25], such as using multiple different instrumental variables for different genes and showing consistent results for the MR analyses with these.

FUTURE PERSPECTIVES

In the first section of this thesis I focused on the discovery of new genetic loci for various inflammatory biomarkers. Even though for each of these markers we substantially increased the number of loci involved in these traits, we still are only able to explain just a fraction of the estimated heritability. As a direct consequence this also limits the capability to provide (causal) functional insights in mechanisms through which genetic variants exert their effects on the phenotype of interest. I will therefore firstly provide some views on how genetic discovery of novel loci can be improved and further extended, followed by a discussion of how recent advances in functional genomics will aid understanding of the link between genotype and phenotype as was aimed for in Section II of this thesis .

9.1. Overcoming the limitations of genetic loci discovery

9.1.1. Increase in power

Super-consortia

As mentioned earlier, one of the simplest ways to uncover substantially larger parts of the so-called 'missing heritability' is by simply increasing sample size, which in turn increases statistical power. From very early on, this has been done through combining GWAS results of individual studies in large meta-analyses, facilitated by the formation of large GWAS consortia, which are often phenotype-specific. Well-known examples are the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC)[26] and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM)[27] consortia that focus on type 2 diabetes and related traits, the Coronary ARtery DIsease Genome wide Replication And Meta-analysis (CARDIoGRAM) consortium[28] that focuses on coronary artery disease. The largest consortium to date is the Genetic Investigation of ANthropometric Traits consortium (GIANT)[29], which focuses on anthropometric traits, of which their first publication on height as a model complex trait, including >180,000 samples[30], showed that their identified genetic loci could approximately explain 10.5% of the explained variance for height. A few years later, a larger effort, of >250,000 samples[31], was able to explain 16%, and estimated that all common variants involved in the phenotype could explain as much as 60%. As various consortia arise working on similar phenotypes, these typically publish either simultaneously ('back to back'), or sometimes efforts would scoop each other. In more recent years, consortia have realised that sample resources are finite, and started to join efforts to further improve sample sizes, such as CARDIoGRAM that has teamed up with the Coronary Artery Disease Genetics Consortium (C4D)[32].

Biobanking

Further efforts to increase sample sizes come from various directions. For example, the recognition of public bodies of the importance of data-driven healthcare, led to the establishment of several biobanks, a well known example being the UK Biobank effort, that aims to improve prevention, diagnosis and treatment of illness, and the promotion of health through society[33]. Up to now, roughly 500,000 samples have been collected with both genetic data and a wide range of phenotypical measurements, which will become available to the wider scientific community in its entirety in the course of 2016. Similar large efforts are Lifelines in the Netherlands[34] and the Estonian Biobank in Estonia[35]. In addition, there are more private biobanks available, such as the The Research Program on Genes, Environment, and Health (RPGEH), a scientific research program from health insurer Kaiser Permanente, that has measured health factors, genetic information from saliva and blood) of also 100K individuals[36], but is planning to reach 500K samples[37]. An interesting source of biobank data comes from 23andMe[38], a privately held personal genomics and biotechnology company based in California. It allows people to send their saliva in a testing kit to have their genomes screened for genes associated with certain inherited conditions, such as cystic fibrosis or sickle cell anaemia, and other genetic markers relating to health and ancestry. Participants can opt to make their genomes available (anonymised) to the scientific community, combined with data retrieved from self-reported health surveys, As of 2015, 23andMe had reached the 1 million mark for genotyped individuals[39], and given that 80% of them opt-in to participate in research[40], this makes it one of the largest medical genetic research databases in the world.

Public availability of data

A final and rather easy way to quickly increase sample sizes or in general making the most of already generated data is simply done by making data publicly available[41]. The Wellcome Trust Case Control Consortium (WTCCC) study on seven common diseases[42] was one of the first that made genotype and phenotype data publicly available to other researchers, making the best use of often publicly funded research. Typical examples are the use of this data for imputation of genotypes, and recently also imputation of epigenomic markers[43] using already generated reference epigenomic datasets, or the inference of eQTLs[44] based only on cohort genotypes and publicly available reference expression data. Imputation of epigenomes is useful as trait-associated variants affect regulatory regions in a cell-type dependent manner[45], and it is not always feasible to map every epigenetic mark in every tissue, cell type and condition of interest. Imputation of epigenomes for specific cell-types enables cell-type / tissue specific postGWAS analyses to understand phenotype specific regulatory consequences of variants identified in a GWAS. Similarly, eQTLs can be cell-type specific[46,47], and appropriate cellular context enables a better understanding of variant consequences on expression level in relevant tissue. This is exactly why platforms such as the database of Genotypes and Phenotypes (dbGaP)[48] or the European Genome Phenome (EGA)[49] archives have been set up. Even though for large meta-GWAS consortia, making data publicly available is difficult, if not impossible due to the many participating studies, all with accompanying release policies and other political stakes, some of these do make the final genome-wide summary statistics available. This allows easy combination of study statistics or replication, and can be done even in the case of unknown proportions of sample overlaps by methods such as METACARPA[50], and also allows much easier and larger-scale Mendelian Randomisation-like analyses based on GWAS summary statistics using packages like Genetics ToolboX (GTX)[51], as used in Chapter 7. To ensure that the vast lakes of generated data are not only publicly available, but also are truly accessible, newer models to improve the portability of data are being developed.[52]

9.1.2. Increase in resolution

Deep sequencing

Regardless of sample size, GWAS analyses are only able to address variants that have been identified and are assessable on a genotype chip, or inferred by imputation. As we continue to sequence many more individuals, more variation is identified, and it has become clear that rarer variation may explain parts of the missing heritability as well for certain traits[53]. In other words, improving genomic resolution is another feasible way to improve genetic discovery, as has already been demonstrated by comparing Hapmap to 1000 Genomes imputation for various traits and the observed improvement in terms of additional loci discovered and fine-mapping[54]. Sequence efforts continue to grow and efforts such as the Haplotype Reference Consortium[55,56], that combine sequencing data from multiple cohorts, provide greatly enriched imputation panels, the creation of which will substantially improve genetic discoveries. Similar efforts to map and collect human genetic variation are ongoing for non-European populations, such as the African Genome Variation Project[57], and efforts by the Sequencing Isolates Consortium (SILC) that brings together a reference panel of sequenced genomes from population isolates. Many more sequence efforts are underway, that may uncover new variants such as the 100,000 Genomes project by Genomics England[58], but also the mind-boggling plan of BGI to sequence 3 million genomes[59], amongst which 1 million human, complemented by projects that will establish deeper and more refined reference maps of the human genome, such as the Human Genome Variation Map[60]. Nevertheless, sequencing such a large number of individuals at reasonable costs is still prohibitive for the majority of GWAS efforts, where at least in the short-term we will remain in an intermediate phase, utilizing whole genome sequences as imputation references and not directly in association analyses[61].

Neglected parts of the genome

Even though many genomes have been sequenced and analysed, certain regions in the genome are largely neglected in genome-wide analyses, such as the Y-chromosome or mitochondrial chromosomes, often due to complications in genotype calling, imputation, and selection of test statistics, as well as a lower proportion of genes, and a lower coverage of current genotyping platforms compared with autosomal chromosomes[62–64]. There is ample evidence for genetic regions in these chromosomes to be involved in a wide range of traits and diseases, including auto-immune diseases[65–67]. Investing in the large-scale analyses of these chromosomes, through efforts such as YGEN, the first international consortium that will assess the influence of Y-chromosome variation on complex traits of public health or evolutionary interest, are likely to provide a multitude of additional insights into the contribution of variants on neglected chromosomes to complex traits[68]. In a similar way, certain genetic regions have been rarely analysed due to the difficulty to type and impute these regions, the best known examples being the immunologically important complex regions such as the HLA region and killer cell immunoglobulin-like receptors (KIRs). Recent methods have however enabled the proper imputation of these regions[69,70], allowing the detailed investigation of their role in human disease.

Fine-mapping

With deeper imputed data, it becomes more likely that a causal variant will be present in the set of analysed variants. It remains however not straightforward to identify these, especially given that the vast majority of associated variants fall outside coding regions[71], whereas until very recent the majority of annotation efforts have focused on coding variants[72]. Newer developments in fine-mapping such as CADD[73] and Eigen[74] have integrated extensive functional genomic annotations into a per-variant score of functionality, including conservation scores, epigenomic annotations and protein-level consequence scores. This importantly enables the prioritisation of variants that are intergenic or in non-coding regions, where previously this was challenging. A well-known example being the 9p21 region for CAD, the first reported CAD locus through GWAS, for which top-variants were intergenic, raising many questions on the biological nature of this association and the responsible causal variant[75]. Other approaches such as CAVIARDB[76], are able to highlight the most likely causal variants in a locus, just using summary statistics and PICS[77], provides similar fine-mapping capabilities, but then also includes prior knowledge such as transcriptional and epigenomic data, promising much improved fine-mapping results.

Capitalising on population diversity

Besides enhancing genetic discoveries by doing the obvious – increasing sample-sizes and analysing enriched sources of variants, one should not underestimate the value of population genetic diversity. Selection pressure due to environmental factors or population bottlenecks may have genetic consequences, that is, genetic variants that are common in one population may be rare or even absent in other populations due to a reduced haplotype diversity, or vice versa. The genetic differences between populations can be exploited in various ways for genetic discovery. Firstly, as also discussed earlier, differences in LD structure may help to further narrow down genetic regions associated with a trait[78], thereby more accurately pinpointing the location of actual causal variants. Secondly, studying different populations can improve the discovery of rare risk variants in loci already highlighted by common variants found by GWAS. When variants rise in frequency, the statistical power to detect these when they underlie the investigated trait will also increase. Whereas in outbred populations certain variants are too low in frequency to be picked up in association studies, higher frequencies of variants in relatively small and homogeneous populations can be identified. This has been demonstrated in, for example, Greek isolates,[79] or the Icelandic population[80]. One caveat however is that differences in variant frequencies and LD-structure between populations implicate that associated variants within one population are not automatically replicable in other populations, hence careful replication strategies must be devised when attempting cross-population replication in GWAS[81].

9.1.3. Increase in throughput and moving towards 'big data'

As discussed earlier, genetic discovery can be greatly aided by increasing sample sizes and high-depth sequencing, which however come at a substantial cost both financially and time wise, though costing much less than the sequencing of the first human genome at $2.7 billion in 1991[82] ($4.7 billion inflation adjusted), whereas now the $1000 dollar genome is a reality[83], or, almost[84]. Newer developments on the horizon do offer prospects to lower these hurdles. One promising technique is nanopore or 'strand sequencing[85,86], a 'lab on a chip' technique that passes a single strand of intact DNA polymers through a protein nanopore, sequencing in real time as the DNA translocates the pore. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane through the other, without the need to amplify DNA, saving expensive reagents. Therefore, the technique is cheaper and has also shown to be much faster[87], and much better at detecting structural variants than previous generation sequencing technologies[88].

Such advances, in concert with ever-increasing biobanking efforts are resulting in huge mountains of data for which matching raw computational power and infrastructure is required. Having the oncoming tidal wave of data in mind, the problem of data processing in genetics should be treated as any other 'big data' issue[89].

On the GWAS analysis front, developments like Genotype Query Tools (GQT)[90], provide novel data indexing strategies and much improved data compression, resulting in substantial query analysis performance improvements over current state-of-the-art tools of over 400 fold. In parallel, algorithms are continuously optimised, whereby borrowing techniques from other fields of research such as signal processing[91] or artificial intelligence[92] offer further improvements in deciphering the human genomic architecture.

On the hardware front other developments continue to contribute to ever more efficient genetic discovery. For example, sequencing pipelines, typically using algorithms running on high-end computer clusters on general CPUs, are being integrated in processors themselves (i.e. 'hard-coded'). Edico Genome's reconfigurable DRAGEN Bio-IT Processor has hard-coded highly optimized algorithms for the full next generation sequencing (NGS) secondary analysis pipeline, including mapping, aligning, sorting, compression and haplotype variant calling, and can be integrated directly into NGS machines and bioinformatics servers[93]. Compared to conventional sequencing pipelines, the company recently showed that this processor reduces the time needed to analyse a whole human genome, from 24 hours to 18 minutes, a speed increase of 80 fold.

Thus, solving data processing problems in genomics requires supercomputing infrastructure and expertise for which industry standards should be adopted. There are two main bottlenecks in the classical way of data management, the first lies in the way files are stored and handled, the second in the way data is processed. Traditionally, files are stored physically on a drive. Disk size then limits the size of the files to be stored. Secondly, when data needs to be processed, data is conventionally moved over a network to be processed by software, which can be extremely slow, in particular for large data sets.

In industry, these problems are tackled by a powerful open source distributed platform to store and manage big data, called Hadoop, which tackles the aforementioned problems in various ways, and holds promise for the application of big data analytics to genomics[94]. What Hadoop does differently, is that it breaks down large datasets into many small files ('chunks' or 'blocks'), which are stored and distributed across the nodes of a computer cluster. Secondly, the MapReduce element of Hadoop, reads data from the database, then puts it into a format suitable for analysis (map), produces customised datasets with only the needed data (reduce), and moves the processing algorithms to the data instead of the other way around[95]. With ever increasing sizes of genomics data, approaches like Hadoop are already being adapted by the scientific community[96,97] , simply out of sheer necessity.

9.2. Increase in understanding (from genotype to phenotype)

As argued in Chapter 8, where we review and evaluate the genetic insights in coronary artery disease, discovery and interpretation of genetic loci influencing traits requires insights from multiple intermediates through which genetic loci exert their effects on the phenotype in order to be able to obtain causal and mechanistic insights. This means integrating genetic data with expression data (eQTL), regulatory elements and effectors, proteomes, metabolomes, and intermediate phenotypes, all of which are interacting biological effector levels that have their effects on the final outcome studied. Yet another phenotypic level to be considered is the microbiome. Apart from more intuitive associations of the microbiome with disease, such as inflammatory bowel disease[98], there are also indications that it influences psychiatric[99,100] and cardiovascular outcomes[101], whereas the composition of the microbiome itself is also influenced by genetic variation of the host[102], which brings yet another dimension to integrative genotype-phenotype analyses. Various efforts have been undertaken to obtain, map and interpret the aforementioned data types. Databases such as Genotype – Tissue Expression (GTEx)[103] and Expression Atlas[104] collect and organise gene expression data derived under different biological conditions and in different tissues, allowing the discovery of eQTL across different tissue types, and combine these sometimes tissue-specific eQTLs with network analyses such as done in Chapter 6.

Similarly, large efforts have been undertaken to map epigenomic data, such as in the ENCODE[105] and Roadmap Epigenetics Project[106], Blueprint Epigenome[107], but also for proteome (HPP)[108], and even microbiome (HMP)[109] data exists on a comparable large scale. Efforts such as HipSci[110], that enable to retrieve these data simultaneously from pluripotent stem cells allow insights in developmental and differentiation mechanisms on a cellular level, and projects such as the UK Digital Heart Project[111], having digitally reconstituted entire human hearts from echocardiographs, may even have the potential to uncover genetic effects on a precise organic scale. Nevertheless, linking and integrating these biological data types will on the one hand undoubtedly improve our understanding of higher-order networks and mechanisms driving inflammatory phenotypes across multiple tissues, but also bring along a great challenge to build statistical models, although many promising methods are already available[112].

9.3. Clinical relevance and impact

Bearing the previous paragraph in mind, translation of identified genetic variation or loci into pathogenic molecular mechanisms appears more feasible than ever before. Nevertheless, the potential clinical relevance remains a key debate. In practise however, GWAS findings have already proven to be clinically informative in a number of ways.

9.3.1. Risk prediction

One of the most straightforward clinical applications of GWAS findings is (genetic) risk prediction of disease, where individual-level risk estimates may help in early intervention and improve diagnostic procedures. Though various methods exist, they make use of the same principle: a set of variants (typically genome-wide) is tested for association with a phenotype, of which a subset that is positively associated is used to produce predicted phenotypes, typically in terms of a continuous genomic risk score (GRS). There have been numerous studies developing risk prediction models using genetic markers, with a few successful examples[113,114], but mostly GWAS-informed risk prediction models have turned out less successful than initially anticipated[113,115]. One of the major reasons being the low effect size of SNPs for common complex diseases identified in GWAS and the relatively low explained variance by these. Therefore, the practical applicability of genetic risk scores in a clinical setting largely depends on the underlying genetic architecture of the disease of interest and achievable sample sizes in studies identifying genetic risk variants[116]. At least in the near future genetic risk scores will, therefore, need to be used as an addition to more established clinical risk factors.

9.3.2. Causal inference and identification of novel risk factors for disease

Given that there is widespread evidence of genetic correlations between traits and diseases[117], there is ample opportunity to investigate causality directions. Using genetic risk scores, it has been confirmed that LDL levels are causal to CAD[118], stressing that LDL-cholesterol lowering interventions such as adjusted diets and LDL lowering medication will have effects, as demonstrated also in a recent meta-analysis of randomised, controlled trials (RCTs)[119]. We ourselves investigated effects of CRP genetic risk scores on 32 different outcomes, and found causal effects of CRP on schizophrenia, notably a protective effect, and nominal effects on blood pressure. Nevertheless, causal findings do not necessarily readily translate to clinical value. In our case for example, it would not make sense to administer medication that increases CRP levels to reduce risk of schizophrenia, thereby increasing the body's inflammatory state, which understandably is a greatly undesired effect. Secondly, schizophrenia is a highly polygenic and complex disease, with a plurality of other factors that may contribute to it's development, of which lowered genetic levels of CRP are one out of many. In the case of our nominal finding of genetically elevated levels of CRP causing elevated levels of blood, administration of anti-inflammatory medication that lowers elevated peripheral CRP may also help to lower blood pressure. The identification of previously unknown risk factors has therefore the potential to aid disease management, but this strongly depends on a number of factors, most importantly the genetic architecture of the disease, and available strategies for intervention.

9.3.3. Disease stratification

Another application comes in the form of disease classification, where for example Sirota and colleagues show that they were able to classify auto-immune diseases[120], identifying SNPs that make an individual susceptible to one class of autoimmune disease whilst simultaneously protecting from diseases in the other autoimmune class. This may enable clinicians to more optimally tailor treatment, as certain drugs for example are known to improve one type of autoimmune disorders, whilst having negative effects on another. A good example for this is infliximab, which is an antibody that binds to TNF, one of the inflammatory markers investigated in this thesis (Chapter 3). Typically prescribed and working well for RA and ankylosing spondylitis (AS)[121,122], it however has no efficacy and sometimes even worsens the condition in individuals with other auto-immune diseases such as multiple sclerosis (MS)[123].

9.3.4. Pharmacogenomics

A major area where GWAS may play a role is in pharmacogenomics. As modern medication only recently appeared as an environmental factor, it will not have caused any negative evolutionary selection pressures on common variants associated with (severe) adverse drug reactions (ADRs). Successful discoveries related to inflammatory disorders include identification of loci for ADRs against Lumiracoxib[124], a drug that is prescribed for the treatment of osteoarthritis and rheumatoid arthritis, causing liver injury, and loci associated with ADRs against thiopurin[125,126], prescribed for autoimmune disorders such as Crohn's disease and rheumatoid arthritis, causing leukopenia and pancreatitis.

Apart from identification of loci related to ADRs, GWAS can also aid in identifying drug targets[127,128]. By making clever use of known gene-drug target databases such as DrugBank[129], therapeutic target database (TTD)[130] and PharmGKB[131], one can overlap genes identified in loci in a meta-analysis, or their interacting gene products[132] and filter out those that appear druggable for further study. As many of the compounds in these databases are already FDA approved, this creates a wealth of opportunities for drug repurposing and repositioning[133], bypassing the necessary lengthy and costly process of clinical safety trials. Lastly, GWAS findings may assist in patient stratification, for example by enabling optimised treatment by tailoring doses of drugs depending on an individuals genetic profile, such as for asthma[134].

In all, genetic discoveries are very likely to contribute in various ways to the realisation of personalised medicine[135], whereby personal (inflammatory) biomarker-specific profiles affected by genetic, clinical and lifestyle factors are likely to play an important role[136].

9.4. Consequences for the general public of advances in genetic discovery

With the promise of personalised medicine, the generation of data lakes, availability of electronic health records (EHRs) for genomic research[137] and these being made available to many scientists, all of which presenting sensitive information regarding individual genomic and health profiles, privacy becomes a key concern. Even though most research bodies have stringent policies regarding anonymisation of data, implemented in law such as the Data Protection Act in the UK[138], if one has access to an individual's DNA, it is rather straightforward to identify presence of these individuals in databases[139]. Just 5000 SNPs are needed in databases of 1000 individuals and the number of SNPs needed declines with the number of samples in a database, even in the presence of sequencing errors and variant-calling differences[140].

Advances in genotyping, making it cheaper and more accessible through efforts such as 23andMe or more widespread use in clinical practice, simultaneously generated interest in the use of genetic information for purposes other than medical research. For example, when linked to electronic health records this information is valuable for insurance or employment purposes, as genetic tests can be used to detect inherited conditions that may have a financial impact. This generates concerns regarding for example employment decisions, possibly the invocation of charging higher insurance premiums or even denial of coverage[141], fears that are most certainly not unreasonable[142–144].

These fears can have various consequences, the most important being individuals that avoid genetic testing, thereby firstly missing out on health benefits these tests can provide, and secondly resulting in negative views on research in medical genomics with all sorts of consequences such as decreased study participation (not giving consent), or even a reduction in funding.

To enable further advances in medical genomic research, it remains therefore pivotal for scientists to properly inform the public about advances and concepts in genomics in an accessible manner, in particular when giving consent to using genetic and phenotypic information in medical genetic studies, and secondly to guarantee absolute anonymity by implementing stringent data management policies[145]. Simultaneously, governments should adhere to extremely strict rules regarding the use of genetic testing, such as in the Netherlands, where the Medical Examination Act greatly restricts the use of genetic information of clients by insurance companies[146].

9.5. Final conclusions and recommendations

To make the most of advances in genetic discovery, I propose the following:

1. Boosting the public availability of scientific output

Firstly, scientific research should be made publicly available whenever possible, especially given the fact that a lot of medical genetic research is more often than not publicly funded. Simultaneously, publication in open access journals should be encouraged. This would have various important benefits, some of the most important being the proper return of investment to the general public, generating the possibility for institutes in countries with challenging economic environments to stay on par with those in more prosperous societies, and last but not least making the most of costly data by recycling these for other experiments. In particular the non-public access to scientific journals, (i.e. these are behind a 'pay-wall') has aggravated many a researcher, so much so that this has culminated in what is known as the 'Pirate Bay' for scientists (http://sci-hub.cc), where 47 million scientific papers are publicly available that otherwise would be behind a pay-wall[147]. Without publicly available data, the majority of work in this thesis would not have been possible, be it the GWAS analyses in Chapters 2 to 5 using Hapmap based data, Chapters 6 using publicly available functional databases, and Chapter 7, in press in an open-access journal (PLOS Medicine) at the time of writing, and using mostly publicly available GWAS summary statistics. Several initiatives are already in place to more or less 'force' scientists, to make data publicly available[148–150], though this is currently more of an exception than a rule.

2. Investment in method development using publicly available data

There should be an even greater extent of method development, to make better use of publicly available summary statistics. Various methods and applications have been mentioned earlier, but some straightforward and widely applicable methods in the context of GWAS meta-analysis are DISTMIX[151] and ImpG[152]. Both are applications that allow the extension of GWAS summary statistics to variants from deeper imputation references using just summary statistics of a GWAS meta-analysis output file without the need of actual re-imputation of individual cohorts to the imputation reference. This firstly greatly reduces computational burden, and secondly can enhance evidence of functional enrichment due to a higher resolution. The re-use of publicly available summary statistics is both financially and time-wise very advantageous in replication studies. A recently developed meta-analysis implementation named METACARPA* allows meta-analysis of summary statistics from public data with one's own study without having to worry about overlap, an initially big hurdle to use publicly available summary statistics for replication.

3. Fostering collaborations between business and science

Thirdly, to fulfil the promise of personalised (genetic) medicine, collaboration between academia and industry is essential[153], and is known to have many other advantages[154], and therefore must be fostered. This does however require improved models of collaboration in order to succeed, overcoming issues such as a mutual lack of trust regarding intellectual property, uncertainty about the potential benefits of working together and outputs from collaborations that are worthwhile for both universities and businesses[155].

4. Adaptation of industrial big data standards

For reasons mentioned earlier in the discussion, (academic) and non-profit medical research entities should aim to adapt industry standards when it comes to big data analysis and management.

5. Giving back to the public

Given the public interest and funding of vast parts of on-going medical genetic research, efforts to communicate with the general public about scientific findings and concepts should be heavily invested in, thereby aiming to create public trust, not the least through warranting absolute privacy when it comes to participation in medical (genetic) studies.

**= https://bitbucket.org/agilly/metacarpa*

REFERENCES

1. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6, 95–108 (2005).
2. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. Am J Hum Genet 90, 7–24 (2012).
3. Benjamin, E. J. et al. Genome-wide association with select biomarker traits in the Framingham Heart Study. BMC Med. Genet. 8 Suppl 1, S11 (2007).
4. GWAS Catalog. The NHGRI-EBI Catalog of published genome-wide association studies (2015). Available at: https://www.ebi.ac.uk/gwas/. (Accessed: 15th January 2016)
5. van der Most, P. J. et al. QCGWAS: A flexible R package for automated quality control of genome-wide association results. Bioinformatics (2014). doi:10.1093/bioinformatics/btt745
6. Ledford, H. Paper on genetics of longevity retracted. Nature News (2011). doi:10.1038/news.2011.429
7. Aruoba, S. B. & Fernández-Villaverde, J. A Comparison of Programming Languages in Economics. (National Bureau of Economic Research, 2014).
8. CRAN Task View: High-Performance and Parallel Computing with R. Available at: https://cran.r-project.org/web/views/HighPerformanceComputing.html. (Accessed: 14th January 2016)
9. ff: memory-efficient storage of large data on disk and fast access functions. Available at: https://cran.r-project.org/web/packages/ff/index.html. (Accessed: 14th January 2016)
10. bigmemory: Manage Massive Matrices with Shared Memory and Memory-Mapped Files. Available at: https://cran.r-project.org/web/packages/bigmemory/index.html. (Accessed: 14th January 2016)
11. Franceschini, N. et al. Discovery and fine mapping of serum protein loci through transethnic meta-analysis. Am. J. Hum. Genet. 91, 744–753 (2012).
12. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164 (2010).
13. Segrè, A. V. et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 6, (2010).
14. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003).
15. Prins, B. P., Lagou, V., Asselbergs, F. W., Snieder, H. & Fu, J. Genetics of coronary artery disease: genome-wide association studies and beyond. Atherosclerosis 225, 1–10 (2012).
16. Vaez, A. et al. In Silico Post Genome-Wide Association Studies Analysis of C-Reactive Protein Loci Suggests an Important Role for Interferons. Circ Cardiovasc Genet 8, 487–497 (2015).
17. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 9 Suppl 1, S4 (2008).
18. Montojo, J. et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 26, 2927–2928 (2010).
19. Emerging Risk Factors Collaboration et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and

mortality: an individual participant meta-analysis. Lancet 375, 132–140 (2010).

20. Emerging Risk Factors Collaboration et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. N. Engl. J. Med. 367, 1310–1320 (2012).

21. Wang, X. et al. Inflammatory markers and risk of type 2 diabetes: a systematic review and meta-analysis. Diabetes Care 36, 166–175 (2013).

22. Henriksen, M. et al. C-reactive protein: a predictive factor and marker of inflammation in inflammatory bowel disease. Results from a prospective population-based study. Gut 57, 1518–1523 (2008).

23. Rhodes, B. et al. A genetic association study of serum acute-phase C-reactive protein levels in rheumatoid arthritis: implications for clinical interpretation. PLoS Med. 7, e1000341 (2010).

24. Ridker, P. M. High-sensitivity C-reactive protein as a predictor of all-cause mortality: implications for research and patient care. Clin. Chem. 54, 234–237 (2008).

25. Glymour, M. M., Tchetgen Tchetgen, E. J. & Robins, J. M. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. Am. J. Epidemiol. 175, 332–339 (2012).

26. MAGIC (the Meta-Analyses of Glucose and Insulin-related traits Consortium). Available at: http://www.magicinvestigators.org. (Accessed: 14th January 2016)

27. DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) consortium. Available at: http://diagram-consortium.org/about.html. (Accessed: 14th January 2016)

28. Preuss, M. et al. Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. Circ Cardiovasc Genet 3, 475–483 (2010).

29. GIANT consortium. Available at: https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium. (Accessed: 14th January 2016)

30. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832–838 (2010).

31. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 46, 1173–1186 (2014).

32. CARDIoGRAMplusC4D Consortium. Available at: http://www.cardiogramplusc4d.org. (Accessed: 14th January 2016)

33. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779 (2015).

34. Scholtens, S. et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int J Epidemiol 44, 1172–1180 (2015).

35. Leitsalu, L. et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol 44, 1137–1147 (2015).

36. Kvale, M. N. et al. Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics genetics.115.178905 (2015). doi:10.1534/genetics.115.178905

37. Schaefer, C. C-A3-04: The Kaiser Permanente Research Program on Genes, Environment and Health: A Resource for Genetic Epidemiology in Adult Health and Aging. Clin Med Res 9, 177–178 (2011).

38. 23andMe. Available at: https://www.23andme.com. (Accessed: 14th January 2016)

39. Power of One Million. 23andMeBlog Available at: https://blog.23andme.com/news/one-in-a-million/. (Accessed: 17th January 2016)

40. 23andMe Therapeutics. 23andMeBlog Available at: https://blog.23andme.com/news/23andme-therapeutics/. (Accessed: 17th January 2016)

41. Singleton, A. B. & Traynor, B. J. For complex disease genetics, collaboration drives progress. Science 347, 1422–1423 (2015).

42. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).

43. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol. 33, 364–376 (2015).

44. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47, 1091–1098 (2015).

45. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet 45, 124–130 (2013).

46. Fu, J. et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 8, e1002431 (2012).

47. Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genet. 9, e1003649 (2013).

48. Mailman, M. D. et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39, 1181–1186 (2007).

49. Lappalainen, I. et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat. Genet. 47, 692–695 (2015).

50. Gilly, Arthur. Metacarpa: META-analysis in C++ Accounting for Relatedness, using arbitrary Precision Arithmetic. Bitbucket (2015). Available at: https://bitbucket.org/agilly/metacarpa. (Accessed: 15th January 2016)

51. Johnson, T. gtx: Genetics ToolboX. (2013).

52. Langille, M. G. I. & Eisen, J. A. BioTorrents: A File Sharing Service for Scientific Data. PLoS One 5, (2010).

53. UK10K Consortium et al. The UK10K project identifies rare variants in health and disease. Nature 526, 82–90 (2015).

54. Wood, A. R. et al. Imputation of Variants from the 1000 Genomes Project Modestly Improves Known Associations and Can Identify Low-frequency Variant - Phenotype Associations Undetected by HapMap Based Imputation. PLoS ONE 8, e64343 (2013).

55. The Haplotype Reference Consortium. Available at: http://www.haplotype-reference-consortium.org. (Accessed: 14th January 2016)

56. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. bioRxiv 035170 (2015). doi:10.1101/035170

57. Gurdasani, D. et al. The African Genome Variation Project shapes medical genetics in Africa. Nature 517, 327–332 (2015).

58. Genomics England is delivering the 100,000 Genomes Project. Available at: http://www.genomicsengland.co.uk. (Accessed: 14th January 2016)

59. BGI Plans to Sequence the World.

Available at: http://biotech.about.com/od/investinginbiotech/a/Bgi-Plans-To-Sequence-The-World.htm. (Accessed: 14th January 2016)

60. UC Santa Cruz to lead effort to build a new map of human genetic variation. UC Santa Cruz to lead effort to build a new map of human genetic variation Available at: http://news.ucsc.edu/2015/01/genome-variation.html. (Accessed: 14th January 2016)

61. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. 47, 1114–1120 (2015).

62. König, I. R., Loley, C., Erdmann, J. & Ziegler, A. How to include chromosome X in your genome-wide association study. Genet. Epidemiol. 38, 97–103 (2014).

63. Wise, A. L., Gyi, L. & Manolio, T. A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am. J. Hum. Genet. 92, 643–647 (2013).

64. Pesole, G. et al. The neglected genome. EMBO Rep 13, 473–474 (2012).

65. Chang, D. et al. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. PLoS ONE 9, e113684 (2014).

66. Case, L. K. et al. Copy number variation in Y chromosome multicopy genes is linked to a paternal parent-of-origin effect on CNS autoimmune disease in female offspring. Genome Biology 16, 28 (2015).

67. Hudson, G., Gomez-Duran, A., Wilson, I. J. & Chinnery, P. F. Recent Mitochondrial DNA Mutations Increase the Risk of Developing Common Late-Onset Human Diseases. PLoS Genet 10, e1004369 (2014).

68. The YGEN Consortium. Available at: https://www.wiki.ed.ac.uk/display/YGEN/Ygen+Home. (Accessed: 15th January 2016)

69. Dilthey, A. T., Moutsianas, L., Leslie, S. & McVean, G. HLA*IMP--an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics 27, 968–972 (2011).

70. Vukcevic, D. et al. Imputation of KIR Types from SNP Variation Data. The American Journal of Human Genetics 97, 593–607 (2015).

71. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. 106, 9362–9367 (2009).

72. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. 12, 628–640 (2011).

73. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310–315 (2014).

74. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet advance online publication, (2016).

75. Samani, N. J. & Schunkert, H. Chromosome 9p21 and Cardiovascular Disease The Story Unfolds. Circ Cardiovasc Genet 1, 81–84 (2008).

76. Chen, W. et al. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. Genetics 200, 719–736 (2015).

77. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).

78. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. Eur J Hum Genet (2016). doi:10.1038/ejhg.2016.1

79. Tachmazidou, I. et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. Nat Commun 4, 2872 (2013).

80. Sulem, P. et al. Identification of low-frequency variants associated with gout and serum uric acid levels. Nat. Genet. 43, 1127–1130 (2011).

81. Senapati, S. et al. Evaluation of European coeliac disease risk variants in a north Indian population. Eur J Hum Genet 23, 530–535 (2015).

82. The Human Genome Project Completion. Available at: https://www.genome.gov/11006943. (Accessed: 14th January 2016)

83. HiSeq X Ten System | 1000 dollar genome sequencing. Available at: http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html. (Accessed: 14th January 2016)

84. Is the $1,000 genome for real? Available at: http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530. (Accessed: 14th January 2016)

85. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. U.S.A. 93, 13770–13773 (1996).

86. Branton, D. et al. The potential and challenges of nanopore sequencing. Nat. Biotechnol. 26, 1146–1153 (2008).

87. USB stick can sequence DNA in seconds. (2012).

88. Tattini, L., D'Aurizio, R. & Magi, A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol 3, (2015).

89. Computational Biology | Profile: Eric Schadt. Nat Biotech 30, 769–770 (2012).

90. Layer, R. M., Kindlon, N., Karczewski, K. J., ExAC, E. A. C. & Quinlan, A. R. Efficient compression and analysis of large genetic variation datasets. bioRxiv (2015). doi:10.1101/018259

91. Kumar, V. et al. Uniform, optimal signal processing of mapped deep-sequencing data. Nat Biotech 31, 615–622 (2013).

92. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. Nat Rev Genet 16, 321–332 (2015).

93. Dragen Bio-IT Platform. Available at: http://www.edicogenome.com/dragen/dragen-bio-it-platform/. (Accessed: 17th January 2016)

94. Chute, C. G. et al. Some experiences and opportunities for big data in translational research. Genet Med 15, 802–809 (2013).

95. Apache Hadoop Fundamentals. Available at: http://www.thegeekstuff.com/2012/01/hadoop-hdfs-mapreduce-intro/. (Accessed: 17th January 2016)

96. Taylor, R. C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics 11 Suppl 12, S1 (2010).

97. Siretskiy, A., Sundqvist, T., Voznesenskiy, M. & Spjuth, O. A quantitative assessment of the Hadoop framework for analyzing massively parallel DNA sequencing data. Gigascience 4, 26 (2015).

98. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory

bowel disease and treatment. Genome Biol. 13, R79 (2012).

99. Hsiao, E. Y. et al. The microbiota modulates gut physiology and behavioral abnormalities associated with autism. Cell 155, 1451–1463 (2013).

100. Castro-Nallar, E. et al. Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. PeerJ 3, e1140 (2015).

101. Wang, Z. et al. Non-lethal Inhibition of Gut Microbial Trimethylamine Production for the Treatment of Atherosclerosis. Cell 163, 1585–1595 (2015).

102. Blekhman, R. et al. Host genetic variation impacts microbiome composition across human body sites. Genome Biol. 16, 191 (2015).

103. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660 (2015).

104. Petryszak, R. et al. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucl. Acids Res. 42, D926–D932 (2014).

105. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

106. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotech 28, 1045–1048 (2010).

107. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. Haematologica 98, 1487–1489 (2013).

108. Legrain, P. et al. The Human Proteome Project: Current State and Future Direction. Mol Cell Proteomics 10, (2011).

109. NIH HMP Working Group et al. The NIH Human Microbiome Project. Genome Res. 19, 2317–2323 (2009).

110. HipSci | Human Induced Pluripotent Stem Cell Initiative. Available at: http://www.hipsci.org. (Accessed: 17th January 2016)

111. UK Digital Heart Project | Using 3D technology to understand heart disease. Available at: http://digital-heart.org. (Accessed: 17th January 2016)

112. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. 16, 85–97 (2015).

113. Jostins, L. & Barrett, J. C. Genetic risk prediction in complex disease. Hum Mol Genet 20, R182–R188 (2011).

114. Clayton, D. G. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet. 5, e1000540 (2009).

115. Schrodi, S. J. et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. Front Genet 5, 162 (2014).

116. Chatterjee, N. et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet 45, 400–405 (2013).

117. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236–1241 (2015).

118. Burgess, S., Freitag, D. F., Khan, H., Gorman, D. N. & Thompson, S. G. Using Multivariable Mendelian Randomization to Disentangle

the Causal Effects of Lipid Fractions. PLoS ONE 9, e108891 (2014).

119. Navarese, E. P. et al. Effects of Proprotein Convertase Subtilisin/Kexin Type 9 Antibodies in Adults With HypercholesterolemiaA Systematic Review and Meta-analysisEffects of PCSK9 Antibodies in Adults With Hypercholesterolemia. Ann Intern Med 163, 40–51 (2015).

120. Sirota, M., Schaub, M. A., Batzoglou, S., Robinson, W. H. & Butte, A. J. Autoimmune Disease Classification by Inverse Association with SNP Alleles. PLoS Genet 5, (2009).

121. Lipsky, P. E. et al. Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. N. Engl. J. Med. 343, 1594–1602 (2000).

122. Grainger, R. & Harrison, A. A. Infliximab in the treatment of ankylosing spondylitis. Biologics 1, 163–171 (2007).

123. Lin, J. et al. TNFalpha blockade in human diseases: an overview of efficacy and safety. Clin. Immunol. 126, 13–30 (2008).

124. Singer, J. B. et al. A genome-wide study identifies HLA alleles associated with lumiracoxib-related liver injury. Nat. Genet. 42, 711–714 (2010).

125. Heap, G. A. et al. HLA-DQA1-HLA-DRB1 variants confer susceptibility to pancreatitis induced by thiopurine immunosuppressants. Nat. Genet. 46, 1131–1134 (2014).

126. Yang, S.-K. et al. A common missense variant in NUDT15 confers susceptibility to thiopurine-induced leukopenia. Nat. Genet. 46, 1017–1020 (2014).

127. Rader, D. J. New Therapies for Coronary Artery Disease: Genetics Provides a Blueprint. Science Translational Medicine 6, 239ps4–239ps4 (2014).

128. Russ, A. P. & Lampel, S. The druggable genome: an update. Drug Discov. Today 10, 1607–1610 (2005).

129. Wishart, D. S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34, D668–672 (2006).

130. Chen, X., Ji, Z. L. & Chen, Y. Z. TTD: Therapeutic Target Database. Nucleic Acids Res. 30, 412–415 (2002).

131. Hewett, M. et al. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res. 30, 163–165 (2002).

132. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. 4, 682–690 (2008).

133. Lussier, Y. A. & Chen, J. L. The Emergence of Genome-Based Drug Repositioning. Sci Transl Med 3, 96ps35 (2011).

134. Wang, Y. et al. Pharmacodynamic genome-wide association study identifies new responsive loci for glucocorticoid intervention in asthma. Pharmacogenomics J 15, 422–429 (2015).

135. Kirchhof, P. et al. The continuum of personalized cardiovascular medicine: a position paper of the European Society of Cardiology. Eur. Heart J. 35, 3250–3257 (2014).

136. Enroth, S., Johansson, A., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. Nat Commun 5, 4684 (2014).

137. Kannry, J. L. & Williams, M. S. Integration of genomics into the electronic health record: mapping terra incognita. Genet Med 15, 757–760 (2013).

138. Data Protection Act 1998. Available at: http://www.legislation.gov.uk/ukpga/1998/29/contents. (Accessed: 17th January 2016)

139. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. Science 339, 321–324 (2013).

140. Shringarpure, S. S. & Bustamante, C. D. Privacy Risks from Genomic Data-Sharing Beacons. The American Journal of Human Genetics 97, 631–646 (2015).

141. Insurance Fears Lead Many to Shun DNA Tests - The New York Times. Available at: http://www.nytimes.com/2008/02/24/health/24dna.html?_r=1. (Accessed: 17th January 2016)

142. Home DNA tests may affect insurance, employment - Health - CBC News. Available at: http://www.cbc.ca/news/health/home-dna-tests-may-affect-insurance-employment-1.3018086. (Accessed: 17th January 2016)

143. 23andMe Is Terrifying, but Not for the Reasons the FDA Thinks - Scientific American. Available at: http://www.scientificamerican.com/article/23andme-is-terrifying-but-not-for-the-reasons-the-fda-thinks/. (Accessed: 17th January 2016)

144. Christiaans, I. et al. Obtaining insurance after DNA diagnostics: a survey among hypertrophic cardiomyopathy mutation carriers. Eur. J. Hum. Genet. 18, 251–253 (2010).

145. Presser, L., Hruskova, M., Rowbottom, H. & Kancir, and J. Care.data and access to UK health records: patient privacy and public trust. Technology Science (2015).

146. Wet van 5 juli 1997, houdende regels tot versterking van de rechts- positie van hen die een medische keuring ondergaan (Wet op de medische keuringen). Staatsblad 365, 1–6 (1997).

147. This pirate website refuses to stop publishing academic papers for free. The Independent (2016). Available at: http://www.independent.co.uk/news/science/pirate-website-offering-millions-of-academic-papers-for-free-refuses-to-close-despite-law-suit-a6875001.html. (Accessed: 15th February 2016)

148. No impact without data access. Nat Genet 47, 691–691 (2015).

149. Barsh, G. S. et al. PLOS Genetics Data Sharing Policy: In Pursuit of Functional Utility. PLoS Genet 11, (2015).

150. the National Institutes of Health Genomic Data Sharing Governance Committees. Data use under the NIH GWAS Data Sharing Policy and future directions. Nat Genet 46, 934–938 (2014).

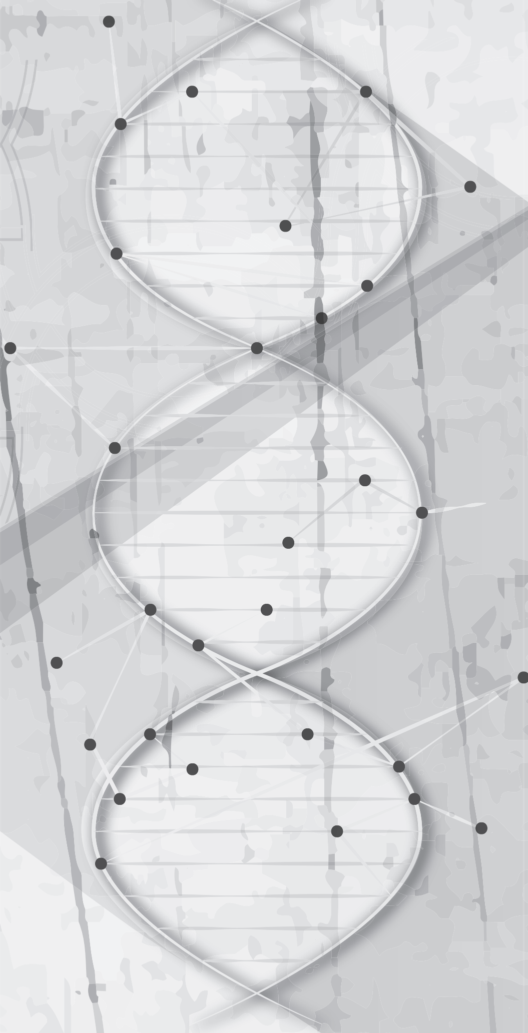
151. Lee, D. et al. DISTMIX: direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. Bioinformatics 31, 3099–3104 (2015).

152. Pasaniuc, B. et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics 30, 2906–2914 (2014).

153. Industrial strength research - Medical Research Council. Available at: https://www.mrc.ac.uk/news/browse/industrial-strength-research/. (Accessed: 18th January 2016)

154. Pronk, J. T. et al. How to set up collaborations between academia and industrial biotech companies. Nat Biotech 33, 237–240 (2015).

155. Jones, S. & Clulow, S. How to foster a culture of collaboration between universities and industry. The Guardian (2012).



SUMMARY

SAMENVATTING

ACKNOWLEDGEMENTS

CURRICULUM VITAE

LIST OF PUBLICATIONS

SUMMARY

Developments in genomics technologies have stood at the basis of the explosion of discovery of genes involved in common traits and diseases with a very complex genetic architecture. A first major leap was the mapping of the entire human genome in 2003, a large global effort taking more than a decade to complete. To give an idea of the size of the human genome; the entire human DNA sequence, consisting in essence of 4 different letters (nucleotides), A,T,C,G, appeared to be 3.2 billion nucleotide bases in length. That is, if you would print every letter of the sequence on paper, the stack of paper would be almost 60 meters high.

Even though our DNA sequence is the same at most places in the genome as compared to other human individuals, in specific positions the nucleotides differ from those of others, referred to as polymorphisms – "the condition of occurring in several different forms". Most typically we focus on changes where at a specific position one single nucleotide has changed, known as a Single Nucleotide Polymorphism (SNP). For example, where the majority of individuals have an "A" at a certain position in the genome, others have a "G". Since all of us have two copies of a chromosome, we can have multiple combinations of these nucleotides: in this case AA, AG, or GG. This combination is called the genotype for a SNP. Many SNPs do not actually have a biological implication, in other words, they are 'neutral', but a fraction of these do have functional consequences when nucleotides change. These polymorphisms can be passed on to the next generation of individuals. In other words, they are heritable, and many of these are known to be involved in complex traits, such as the serum levels of inflammatory biomarkers studied in this thesis.

Since analysis of all possible SNPs for many individuals is not financially feasible, oftentimes a selection of SNPs is made for analyses. Typically we make use of so-called tag SNPS, which are SNPs that are highly correlated with neighbouring SNPS (this is called a haplotype). This correlation is called the linkage disequilibrium correlation (LD r^2). In that way, we only have to obtain the genotypes for a fairly limited set of tag SNPs instead of all SNPs throughout the genome, reducing analysis costs. Making use of the same principle of the LD structure of the genome, we can then infer the state of the correlated SNPs in individuals even if we did not measure them, by using the known haplotypes from a reference panel. This is known as imputation.

To find out if our SNPs are involved in a quantitative trait of interest, we perform statistical tests, in this case association analyses. They allow us to find out if our SNP, such as a change of A to G (see above), is accompanied with a change in our trait of interest. In this thesis it often concerned levels of inflammatory biomarkers. Performing these statistical analyses for SNPs across the entire genome is referred to as a genome-wide association study (GWAS). If we measure the effect for each additional change of a nucleotide (or 'allele'), for the genotype of a certain SNP (i.e. the effect of 0,1 or 2 G alleles in the AA, AG, GG genotypes in the example above), we thereby make use of a so-called 'additive' model. This allows us to discover regions in the genome ('genetic loci') that sometimes contain genes that are involved in our trait of interest.
The genetic architecture of the traits we study is complex – there are typically many genetic loci in the genome involved, and their individual influence is small, which makes it difficult to find them. More technically speaking, we need a lot of statistical 'power' to find the genetic loci involved. The most straightforward way is increasing the number of individuals that is used in the analyses, by combining the GWAS analyses from many studies, referred to as meta-analyses.

Most of the analyses in this thesis, are either meta-analyses of GWAS studies ('meta-GWAS'), or have made use of the results of these. Meta-analyses are combinations of results from different GWAS studies for the same trait, where typically these are results for > 2.5 million SNPs, with different variables that inform us about the statistical results for these SNPs, basic information such as the position in the genome, and the quality of these SNPs. As the data originates from many different collaborators that have different platforms to measure traits of interest and different so-called pipelines for the analysis of the data concerning millions of data points, it is easy to introduce mistakes in the results, which are difficult to check and correct manually.

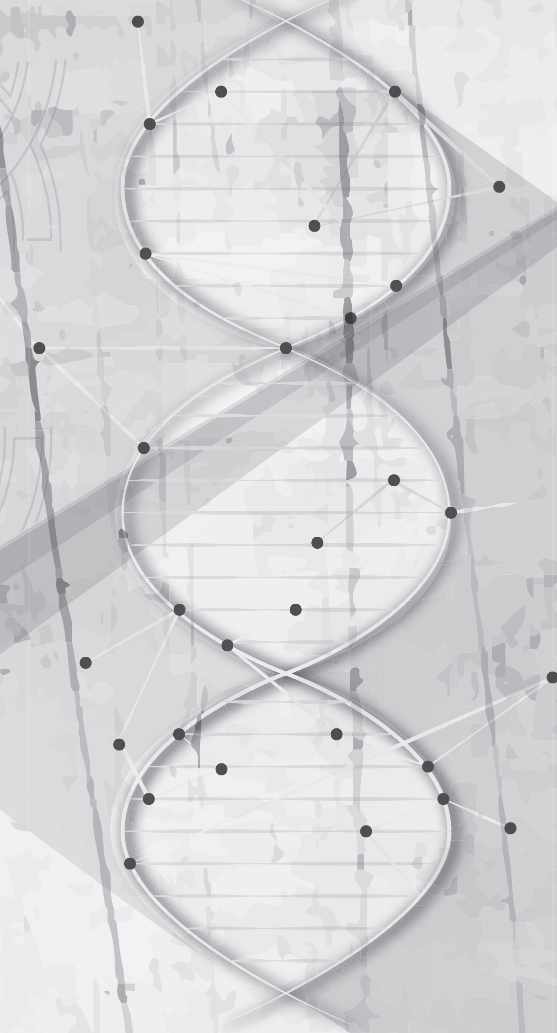
Therefore, in Chapter 2, we first developed a software pipeline, in the form of a statistical software package named QCGWAS, which can automatically perform quality control ("QC") of the GWAS summary statistics files, and standardise these to a format suitable for meta-analysis. We used this pipeline subsequently for the GWAS meta-analyses in Chapters 3 and 4 for two well-known inflammatory biomarkers, Tumor Necrosis Factor Alpha (TNF-α) and Interleukin-6, both of which are known to be elevated in various diseases, ranging from auto-immune diseases such as rheumatoid arthritis to cardiovascular diseases, such as coronary artery disease. In these two chapters, we identified six genetic loci (most of them new) for these markers. These loci contained various inflammation-related genes and SNPs with functional biological consequences in some instances. We also showed that the loci we identified are involved in a number of various diseases, giving initial evidence for potential (partially) shared genetic architecture. This is also referred to as pleiotropy – when genes or genetic regions are involved in multiple traits.

In most GWAS analyses it is difficult to pinpoint the exact gene in a locus that is causally involved, let alone the causal SNP, both due to the strong LD correlation between SNPs in certain regions. However, LD structures differ between populations, and by actually comparing GWAS results from a trait between two populations and making use of LD-structure differences, we can actually narrow down the region that contains the causal SNP, and thereby increase our chances to pinpoint it. This is exactly what we have done in Chapter 5, where we sought to identify genetic loci that influence serum levels of total protein and albumin in the blood, two biomarkers of which changes in concentration are associated with various diseases. We analysed data from two populations; Europeans and Japanese, and obtained substantial improvements in the resolution of fine mapping of potential causal variants by leveraging the ethnic differences in the distribution of LD between these. One of the best improvements in terms of resolution occurred in the 6q21.3 locus that we identified for total protein, where in the European-only analysis a set of 14 SNPs could contain the causal SNP with 99% certainty, but after combining the results with those done on Japanese ancestry, this was reduced to just 3 SNPs.

In both Chapter 5 and Chapter 6, we also show the importance to use other types of molecular 'in-silico' data to better understand the genetic loci that we identified. In Chapter 5, we made some initial steps, by performing pathway and protein interaction network analyses to see which molecular pathways are affected by our genetic loci, and to see what other proteins the products of genes in our loci (also proteins) interact with. In Chapter 6, we formalised a pipeline for these types of analyses, also referred to as 'post-GWAS' analyses, and used this to identify previously unknown mechanisms involved in determining levels of C-Reactive Protein (CRP), such as interferon-related pathways. Importantly, we also show that quite a few of our SNPs have an effect on gene expression.

CRP is one of the most widely used inflammatory biomarkers in a clinical setting, and typically is used to measure the general degree of inflammation in various conditions, ranging from cardiovascular, neuropsychiatric to auto-immune disease. One long-standing question regarding CRP has been whether it is a consequence of disease, or whether it actually is causally involved in disease development and progression. Genetic loci identified through GWAS can actually aid answering that question. In practice, it is difficult to obtain data for large numbers of patients with different diseases and their measured CRP levels. However, instead we can use the genetic variants that we know influence CRP levels, and measure their cumulative effects in a case versus control GWAS for a certain disease, and perform a statistical test to check whether their effect is significant. In other words, we don't measure the effects of CRP on disease directly, but through the genetic variants that influence CRP levels, just using GWAS summary statistics that are made available by other scientists. Using this technique, we were able to demonstrate that for most common complex diseases CRP is not causally involved. The only exception was schizophrenia for which we showed a potentially causal protective involvement of CRP: when our CRP SNPs predict that CRP levels go up, the risk to develop schizophrenia goes down.

In Chapter 8, I argue that performing GWAS analyses for traits, here with a focus on Coronary Artery Disease, is just a first step towards a molecular understanding of a trait, and that in the end the most promising way forward is to integrate multiple layers of data from molecular intermediates that act together in certain pathways in order to complete the picture.



SUMMARY

SAMENVATTING

ACKNOWLEDGEMENTS

CURRICULUM VITAE

LIST OF PUBLICATIONS

SAMENVATTING

Ontwikkelingen in genomics technologieën stonden aan de basis van de explosie van de ontdekking van genen die betrokken zijn bij veelvoorkomende ziekten en eigenschappen met een (zeer) complexe genetische architectuur. Een eerste grote stap was het in kaart brengen van het gehele humane genoom in 2003; een wereldwijde inspanning die meer dan een decennium in beslag nam. Het humane genoom is groot; de gehele menselijke DNA sequentie, in principe bestaande uit 4 letters (nucleotiden), A,T,C,G, bleek 3.2 miljard nucleotide basen lang te zijn. Als je elke letter van de sequentie zou uitprinten op papier, dan zou de stapel van al deze vellen bijna 60 meter hoog zijn.

Ondanks dat onze DNA sequentie hetzelfde is op de meeste plekken in het genoom vergeleken met andere mensen, verschillen de nucleotiden op specifieke posities tussen individuen, dit worden ook wel polymorfismen genoemd – “het voorkomen in verschillende vormen”. Meestal focussen we op veranderingen waarbij op een specifieke positie één enkele nucleotide is veranderd, een Single Nucleotide Polypmorphism (SNP) genoemd. In het Nederlands zou je het kunnen vertalen naar een “Enkelvoudig Nucleotide Polymorfisme”. Bijvoorbeeld: waar de meeste individuen een “A” hebben op een bepaalde positie in het genoom, hebben anderen een “G”. Omdat iedereen twee kopieën heeft van een chromosoom, kunnen we verschillende combinaties hebben van deze nucleotiden; in dit geval AA, AG, of GG. Deze combinaties worden de genotypen voor een SNP genoemd. De meeste SNPs hebben geen biologische gevolgen, met andere woorden, ze zijn ‘neutraal’, maar een fractie ervan heeft wel degelijk functionele gevolgen wanneer de nucleotide verandert. Deze polymorfismen kunnen worden doorgegeven naar de volgende generatie van individuen. Anders gezegd, ze zijn erfelijk, en velen zijn betrokken bij het bepalen van complexe (biologische) eigenschappen, zoals de niveaus van de ontstekingseiwitten in het bloed die in dit proefschrift bestudeerd zijn.

Aangezien het analyseren van alle mogelijke SNPs voor een grote studie met veel participanten financieel niet haalbaar is, wordt er vaak een selectie van SNPs gemaakt voor analyses. Typisch wordt er gebruik gemaakt van zogenaamde ‘tag SNPS’, dat zijn SNPs die sterk gecorreleerd zijn met de naburige SNPS (ook wel een haplotype genoemd). Deze correlatie wordt de linkage disequilibrium correlatie genoemd (LD r^2). Op deze manier hebben we alleen de genotypen nodig voor een redelijk beperkt aantal tag SNPs in plaats van alle SNPs in het hele genoom, wat de analysekosten vermindert. Door op dezelfde manier gebruik te maken van het principe van de LD structuur in het genoom en bekende haplotypen van referentiegenomen, kunnen we de genotypen van SNPs afleiden door middel van gecorreleerde SNPs, zelfs als we deze SNPs niet direct hebben gemeten in deze personen. Dit wordt imputatie genoemd.

Om erachter te komen of deze SNPs betrokken zijn bij een kwantitatieve biologische eigenschap waarin we geïnteresseerd zijn, maken we gebruik van statistische berekeningen, in dit geval associatie analyses. Deze stellen ons in staat om te bepalen of een bepaalde SNP zoals een verandering van een A in een G (zie boven), gepaard gaat met een verandering in de biologische eigenschap die we bestuderen. In dit proefschrift ging het om bloedniveaus van ontstekingseiwitten. Het uitvoeren van deze statistische analyses voor SNPs over het gehele genoom wordt een genoombrede associatie studie (GWAS) genoemd. Als we het effect meten voor elke toegevoegde verandering van een nucleotide (of 'allel'), voor het genotype van een bepaalde SNP (dwz het effect van 0, 1 of 2 G allelen in de AA, AG of GG genotypen van het voorbeeld hierboven), maken we daarbij gebruik van een zgn. 'additief' model. Dit stelt ons in staat om gebieden in het genoom ('genetische loci') te ontdekken waarin zich soms genen bevinden die betrokken zijn bij onze bestudeerde eigenschappen.

De genetische architectuur van de eigenschappen die we bestuderen is complex - gewoonlijk zijn er vele genetische loci in het genoom bij betrokken en hun individuele invloed is klein, waardoor het moeilijk is om ze te ontdekken. Meer technisch uitgedrukt, kunnen we stellen dat we zeer veel statistische 'power' nodig hebben om de genetische loci te ontdekken die betrokken zijn bij complexe biologische eigenschappen. De meest eenvoudige manier om dit te bereiken is door meer individuen te includeren door de resultaten van verschillende GWAS analyses in meerdere cohorten te combineren. Dit wordt aangeduid als meta-analyses.

Het merendeel van de analyses in dit proefschrift, zijn meta-analyses van GWAS studies ('meta-GWAS'). In andere gevallen maken we gebruik van de meta-analyse resultaten van andere studies die publiek toegankelijk zijn gemaakt. Meta-analyses, zoals eerder beschreven, zijn combinaties van resultaten van verschillende individuele GWAS studies voor dezelfde bestudeerde biologische eigenschap. Deze omvatten gewoonlijk resultaten voor > 2,5 miljoen SNPs gespecificeerd door verschillende variabelen, zoals de resultaten van de statistische berekeningen, de positie in het genoom en de kwaliteit van de SNP. Omdat de gegevens afkomstig zijn van verschillende analisten en groepen, die elk verschillende methoden gebruiken om de bestudeerde biologische eigenschappen te meten en gebruik maken van verschillende zogenaamde 'pijplijnen' voor de analyses van de gegevens die miljoenen datapunten omvat, is het gemakkelijk om fouten te maken in de resultaten. Deze fouten zijn lastig te controleren, laat staan dat het mogelijk is om deze allemaal handmatig te corrigeren.

Dat is precies de reden waarom we in hoofdstuk 2, eerst een 'pijplijn', ontwikkelden in de vorm van een statistisch softwarepakket genaamd QCGWAS, dat automatisch kwaliteitscontroles ("QC" – "Quality Control") kan uitvoeren op de GWAS resultaatsbestanden en deze standaardiseert naar een formaat geschikt voor meta-analyses. We hebben deze 'pijplijn' vervolgens gebruikt voor de GWAS meta-analyses in hoofdstukken 3 en 4 voor twee bekende ontstekingseiwitten, Tumor Necrose Factor alfa (TNF-α) en Interleukine-6 (IL-6). Beide zijn bekend vanwege verhoogde niveaus bij verschillende ziekten, variërend van auto-immuunziekten zoals reumatoïde artritis tot cardiovasculaire aandoeningen, zoals coronaire hartziekte. In deze twee hoofdstukken identificeerden we zes genetische loci (de meeste voorheen onbekend) voor deze ontstekingseiwitten. Deze loci omvatten verschillende ontstekings-gerelateerde genen en in sommige gevallen SNPs met functionele biologische gevolgen. We toonden ook aan dat de loci die we identificeerden betrokken zijn bij een aantal verschillende ziekten, met initiële aanwijzingen voor een potentiële gedeelde genetische architectuur. Dit wordt ook wel pleiotropie genoemd – het verschijnsel dat genen of genetische gebieden invloed hebben op meerdere biologische eigenschappen.

In de meeste GWAS analyses is het moeilijk om het exacte gen te lokaliseren in een genetisch locus dat causaal betrokken is bij de onderzochte eigenschap, laat staan de causale SNP, beide vanwege de sterke LD correlatie tussen SNPs in bepaalde gebieden. LD structuren verschillen echter tussen populaties. Juist door de GWAS resultaten tussen twee etnische populaties voor bepaalde biologische eigenschap te vergelijken en gebruik te maken van verschillen in LD-structuur, kunnen we het gebied in het genoom dat de daadwerkelijke causale SNP bevat verkleinen, waardoor de kansen stijgen om deze te ontdekken. Dit is precies wat we in hoofdstuk 5 hebben gedaan, waar we trachtten om genetische loci te ontdekken die bloedniveaus beïnvloeden van totaal eiwit en albumine, twee biomarkers waarvan veranderingen in concentraties geassocieerd worden met diverse ziekten. We analyseerden de gegevens van twee populaties; uit Europa en Japan, en behaalden aanzienlijke verbeteringen in resolutie bij het in kaart brengen van mogelijke causale varianten door gebruik te maken van de etnische verschillen in LD structuur. Een van de meest spraakmakende verbeteringen in termen van resolutie vonden we in het 6q21.3 gebied dat we ontdekten voor totale eiwit concentratie. In de analyse met individuen uit alleen Europa vonden we met 99% zekerheid dat de causale SNP één van 14 gevonden SNPs zou kunnen zijn, echter na het combineren van de resultaten met de analyse voor individuen van Japanse afkomst, werd dit aantal teruggebracht tot slechts 3 SNPs.

Zowel in hoofdstuk 5 als hoofdstuk 6 laten we het belang zien van het gebruik van andere soorten moleculaire 'in-silico 'data om genetische gebieden die we hebben ontdekt beter te kunnen begrijpen. In hoofdstuk 5, zetten we een aantal initiële stappen in dit type onderzoek, door het analyseren van moleculaire routes en eiwit-interactie netwerkanalyses om te zien welke moleculaire processen worden beïnvloed door onze genetische gebieden en uit te vinden met welke andere eiwitten de producten van de genen in onze gebieden (ook eiwitten) interacteren. In hoofdstuk 6 hebben we dit soort analyses, ook wel bekend als 'post-GWAS' analyses, formeel geïntegreerd in een 'pijplijn'. Met behulp hiervan ontdekten we voorheen onbekende mechanismen die betrokken zijn bij het bepalen van niveaus van C-Reactive Protein (CRP), zoals interferon-gerelateerde moleculaire processen. Wat minstens zo belangrijk is, is dat we aantoonden dat ook dat een substantiële proportie van onze SNPs een effect hebben op genexpressie.

CRP is een van de meest gebruikte ontstekingseiwitten in een klinische context, en wordt normaal gesproken benut om de mate van ontsteking in verschillende omstandigheden en aandoeningen te meten, variërend van cardiovasculaire tot neuro-psychiatrische en auto-immuun aandoeningen. Een vraag die onderzoekers al lang bezighoudt is of een verhoogd CRP niveau het gevolg is van een aandoening, of dat het daadwerkelijk oorzakelijk betrokken is bij de ontwikkeling en progressie van bepaalde ziekten. Genetische gebieden ontdekt door middel van GWAS kunnen erbij helpen om deze vraag te beantwoorden. In de praktijk is het moeilijk om gegevens voor een groot aantal patiënten met verschillende ziekten én gemeten CRP niveaus te verzamelen. We kunnen echter in plaats daarvan de genetische varianten gebruiken waarvan we weten dat ze CRP niveaus beïnvloeden, en hun cumulatieve effecten bepalen uit een GWAS voor een bepaalde ziekte, gevolgd door het uitvoeren van een andere statistische test om te kijken of hun gezamenlijk effect een significante invloed heeft op de ziekte in kwestie. Met andere woorden, wat we hebben gedaan is niet direct het effect van CRP op een ziekte gemeten, maar in plaats daarvan hebben we de genetische varianten gebruikt die CRP niveaus beïnvloeden, slechts met behulp van GWAS resultaten die door andere wetenschappers ter beschikking werden gesteld (dus geen individuele data). Door gebruik te maken van deze techniek konden we aantonen dat CRP niet causaal betrokken is bij de meeste veel voorkomende complexe ziekten. De enige uitzondering was schizofrenie, waarvoor wij een mogelijke causale beschermende betrokkenheid van CRP aantoonden: wanneer onze CRP SNPs voorspellen dat CRP levels omhoog gaan, dan gaat risico voor het ontwikkelen van schizofrenie omlaag.

In hoofdstuk 8, betoog ik dat het uitvoeren van GWAS analyses voor biologische eigenschappen, in dit geval met specifieke focus op coronaire hartziekte, slechts een eerste stap is op weg naar een beter moleculair begrip van een ziekte of eigenschap. Uiteindelijk de meest veelbelovende manier om het plaatje compleet te maken is door het integreren van verschillende lagen van data over moleculaire tussenproducten, die met elkaar interacteren in onderliggende fysiologische 'pathways'.